polynomial regression

Nonlinear regression

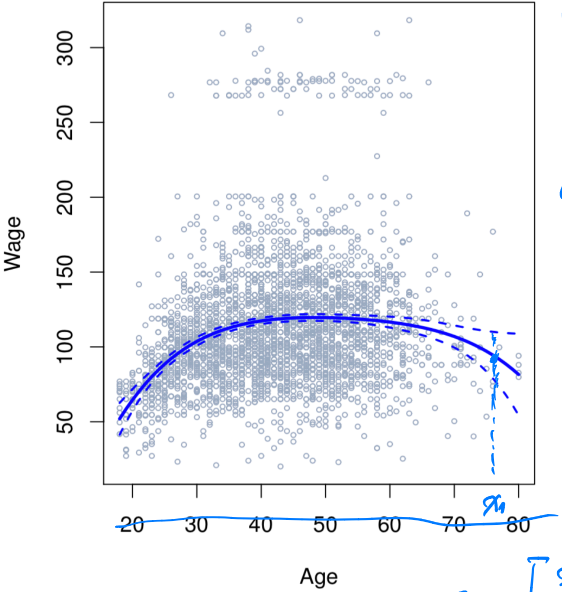
$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_d x_i^d$, $\underline{d\text{-degree polynomial}}$, $df = d+1$.

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_d \end{bmatrix} \quad \underline{\text{OLS}} \quad \underline{X}, Y$$

$X_1 = X$, $X_2 = X^2$, $\cdots$ $X_d = X^d$, $X' = [X_1, \cdots X_d]_{n \times d}$

$$\hat{\beta} = \operatorname{argmin} \underline{\sum_{i=1}^{n}} \underline{(y_i - \beta^T x_i)} \quad \text{RSS}$$

$$\underline{\hat{\beta} = (X'^T X')^{-1} X'^T} Y.$$



point estimation:

$$\hat{f}(x_0) = \hat{\beta}_0 + \cdots + \hat{\beta}_d x_i^d$$

interval estimation:

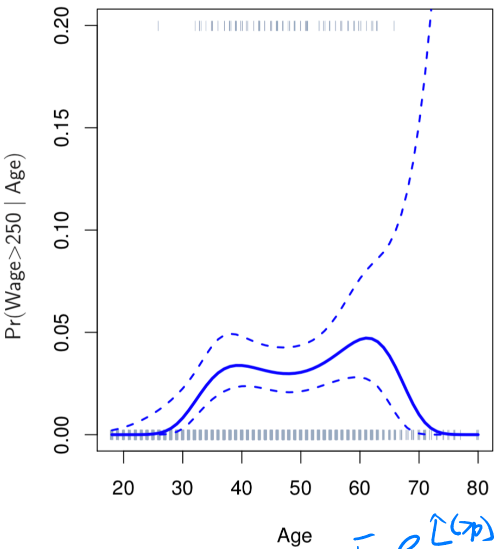
95% CI, $z_{\frac{\alpha}{2}} = 1.96$.

$$\hat{f}(x_0) \pm 1.96\, se(\hat{f}(x_0))$$

$$= \sqrt{Var(\hat{f}(x_0))}$$

$$= \sqrt{Z^T (X^T X)^{-1} Z}.$$

$$Z = \begin{bmatrix} x_0 \\ \vdots \\ x_0^d \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1 & \cdots & x_1^d \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^d \end{bmatrix}$$



$f(x_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i + \cdots + \beta_d x_i^d$

MLE 求 $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_d \end{bmatrix}$

$\hat{p}_i = \frac{e^{\hat{f}(x_0)}}{1+e^{\hat{f}(x_0)}}$, $\hat{f}(x_0) = \hat{\beta}_0 + \cdots + \hat{\beta}_d x_i^d$

95% CI of $\hat{f}(x_0)$ 算出

$[\hat{L}(x_0), \hat{U}(x_0)]$

$\hat{f}(x_0) \pm 1.96\sqrt{Var(\hat{f}(x_0))}$

$\hat{p}_0 \in \left[\frac{e^{\hat{L}(x_0)}}{1+e^{\hat{L}(x_0)}}, \frac{e^{\hat{U}(x_0)}}{1+e^{\hat{U}(x_0)}}\right]$ GLM delta method

① $d \to$ test MSE

Tuning parameter to choose optimal d

② Hypothesis test

ANOVA

```
fit.1=lm(wage~age,data=Wage)
fit.2=lm(wage~poly(age,2),data=Wage)
fit.3=lm(wage~poly(age,3),data=Wage)
fit.4=lm(wage~poly(age,4),data=Wage)
fit.5=lm(wage~poly(age,5),data=Wage)
anova(fit.1,fit.2,fit.3,fit.4,fit.5)
```

$y_i = \beta_0 + \beta_1 x_i$

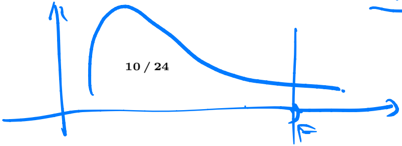
$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$

$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$

```
## Analysis of Variance Table
##
## Model 1: wage ~ age
## Model 2: wage ~ poly(age, 2)
## Model 3: wage ~ poly(age, 3)
## Model 4: wage ~ poly(age, 4)
## Model 5: wage ~ poly(age, 5)
##   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
## 1   2998 5022216
## 2   2997 4793430  1    228786 143.5931 < 2.2e-16 ***
## 3   2996 4777674  1     15756   9.8888  0.001679 **
## 4   2995 4771604  1      6070   3.8098  0.051046 .
## 5   2994 4770322  1      1283   0.8050  0.369682
## ---
```

$H_0: \beta_3 = 0$ vs $H_a: \beta_3 \neq 0$

$F_{1,2996} = \dfrac{\sigma_{\beta_3}}{MSE_2 = RSS/df_{residual}}$



Orthogonal polynomial regression

lm(y~poly(x,4, raw=T))

$$y_i = \beta_0 + \beta_1 x_i + \dots + \beta_d x_i^d + \varepsilon_i$$

$$\Phi = \begin{bmatrix} \Phi_1 \\ \vdots \\ \Phi_m \end{bmatrix}, \text{ such that } \underline{z_m} = \sum_{i=1}^{p} \phi_{mj} x^j, \quad \frac{z_1, z_2 \dots z_m}{z_i \cdot z_j^T = 0}$$

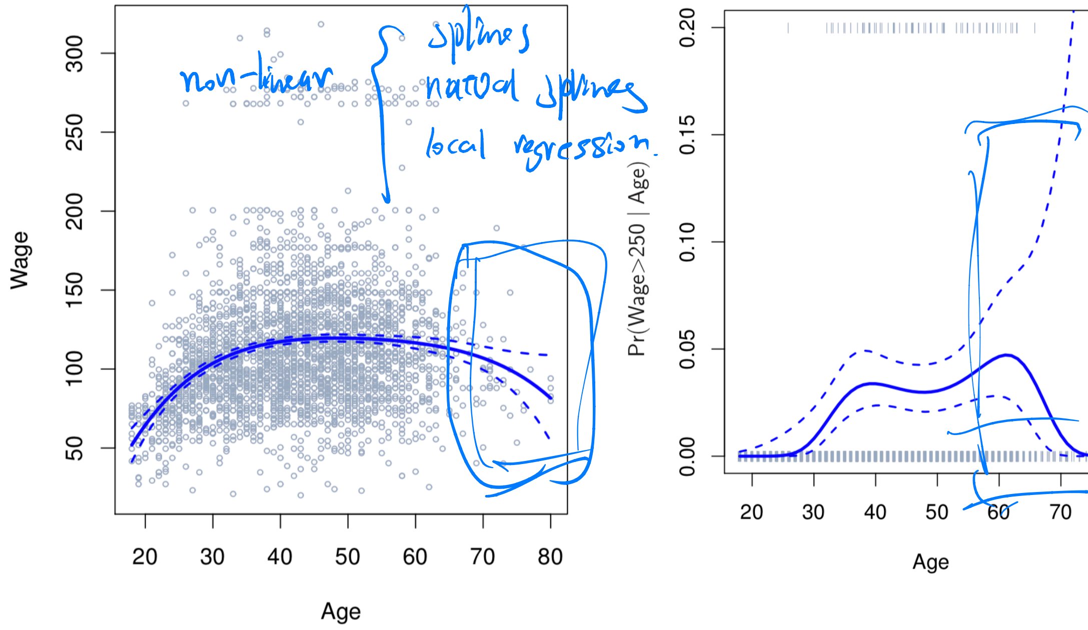
$$y_i = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + \dots + \theta_d x_i^d + \varepsilon_i$$

Orthogonal polynomial regression

t-test

```r
coef(summary(fit.5))
```

```
##                  Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)     111.70361  0.7287647 153.2780243 0.000000e+00
## poly(age, 5)1   447.06785 39.9160847  11.2001930 1.491111e-28
## poly(age, 5)2  -478.31581 39.9160847 -11.9830341 2.367734e-32
## poly(age, 5)3   125.52169 39.9160847   3.1446392 1.679213e-03
## poly(age, 5)4   -77.91118 39.9160847  -1.9518743 5.104623e-02
## poly(age, 5)5   -35.81289 39.9160847  -0.8972045 3.696820e-01
```



Wage
Age
Pr(Wage>250 | Age)
Age
non-linear
splines
natural splines
local regression.