20) Polynomial Regression, Step Functions, Basis Functions

Vitor Kamada

March 2018

Reference

Tables, Graphics, and Figures from

An Introduction to Statistical Learning

James et al. (2017): Chapters: 7.1, 7.2, 7.3, and 7.8.1

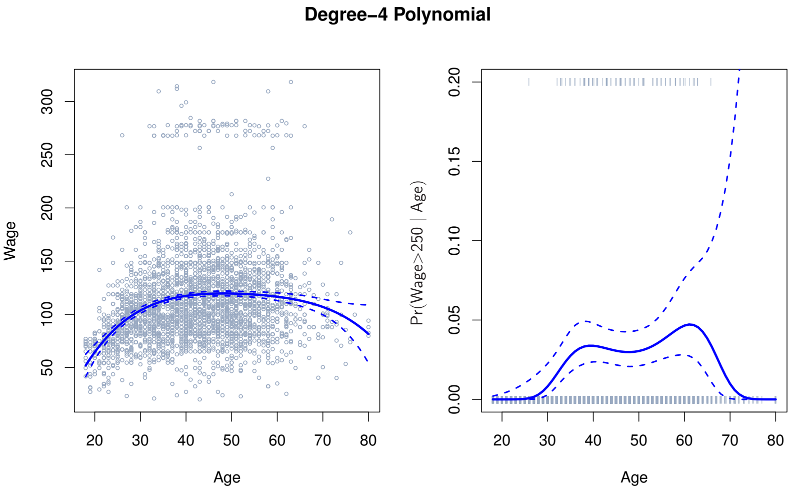
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d + \epsilon_i$$

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

$$Pr(y_i > 250|x_i)$$

$$= \frac{exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}{1 + exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}$$

The Wage Data: Males in the Atlantic Region of the United States



Step Functions

$$C_0(X) = I(X < c_1)$$

$$C_1(X) = I(c_1 \leq X < c_2)$$

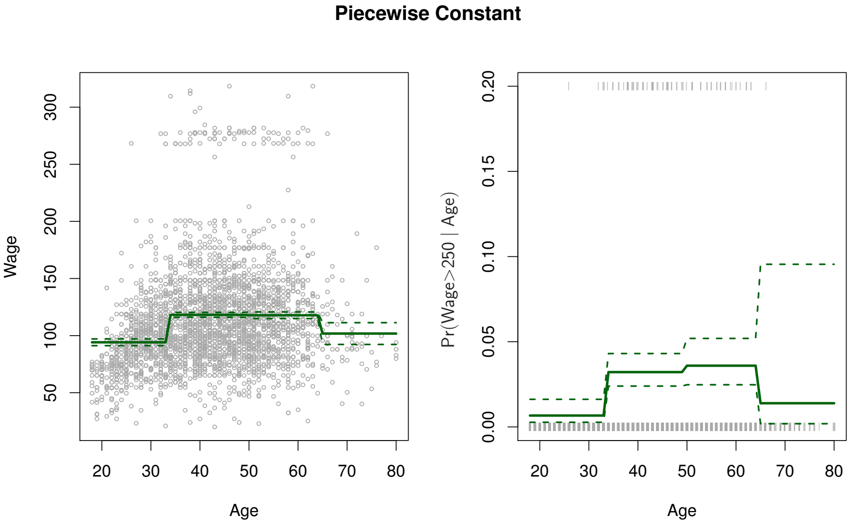
$$C_2(X) = I(c_2 \leq X < c_3)$$

$$C_K(X) = I(c_K \leq X)$$

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \ldots + \beta_K C_K(x_i) + \epsilon_i$$

$$Pr(y_i > 250|x_i) = \frac{exp(\beta_0+\beta_1 C_1(x_i)+\ldots+\beta_K C_K(x_i))}{1+exp(\beta_0+\beta_1 C_1(x_i)+\ldots+\beta_K C_K(x_i))}$$

Step Functions: OLS and Logit



Basis Functions

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + ... + \beta_K b_K(x_i) + \epsilon_i$$

$$b_j(x_i) = x_i^j$$

$$b_j(x_i) = I(c_j \leq x_i < c_{j+1})$$

fit=lm(wage~poly(age,4),data=Wage)

coef(summary(fit))

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	111.70361	0.7287409	153.283015	0.000000e+00
poly(age, 4)1	447.06785	39.9147851	11.200558	1.484604e-28
poly(age, 4)2	-478.31581	39.9147851	-11.983424	2.355831e-32
poly(age, 4)3	125.52169	39.9147851	3.144742	1.678622e-03
poly(age, 4)4	-77.91118	39.9147851	-1.951938	5.103865e-02

fit2=lm(wage~poly(age,4,raw=T),data=Wage)

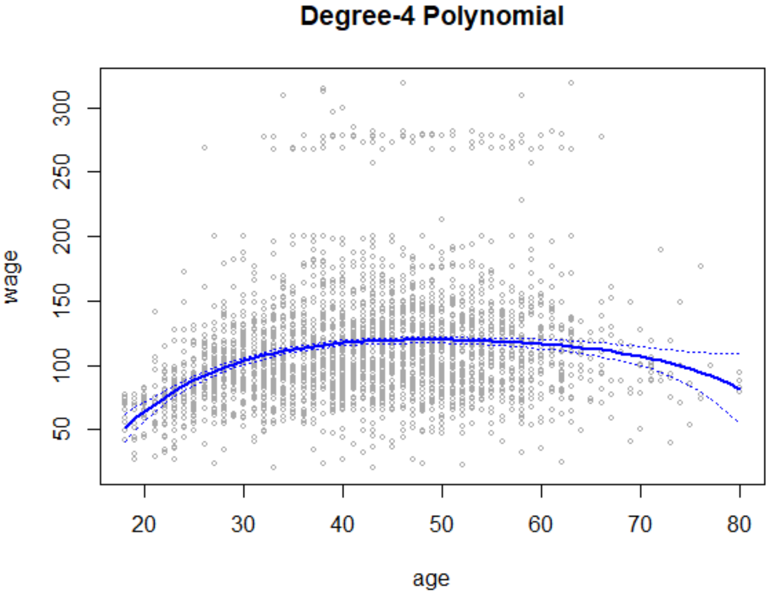
coef(summary(fit2))

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.841542e+02	6.004038e+01	-3.067172	0.0021802539
poly(age, 4, raw = T)1	2.124552e+01	5.886748e+00	3.609042	0.0003123618
poly(age, 4, raw = T)2	-5.638593e-01	2.061083e-01	-2.735743	0.0062606446
poly(age, 4, raw = T)3	6.810688e-03	3.065931e-03	2.221409	0.0263977518
poly(age, 4, raw = T)4	-3.203830e-05	1.641359e-05	-1.951938	0.0510386498

agelims=range(age)

```
agelims=range(age)
age.grid=seq(from=agelims[1],to=agelims[2])
preds=predict(fit,newdata=list(age=age.grid),se=TRUE)
se.bands=cbind(preds$fit+2*preds$se.fit,
preds$fit-2*preds$se.fit)
plot(age,wage,xlim=agelims,cex=.5,col="darkgrey")
title("Degree-4 Polynomial",outer=F)
lines(age.grid,preds$fit,lwd=2,col="blue")
matlines(age.grid,se.bands,lwd=1,col="blue",lty=3)
```

Degree-4 Polynomial



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)
```

```
  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    
```

fit=glm(I(wage>250)~poly(age,4), data=Wage, family=binomial)

preds=predict(fit,newdata=list(age=age.grid),se=T)

$$log\left(\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)}\right) = X\hat{\beta}$$

pfit=exp(preds$fit)/(1+exp(preds$fit))

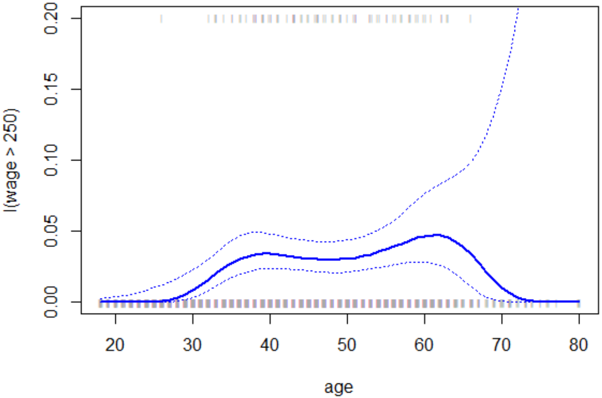
$$Pr(Y = 1|X) = \frac{exp(X\hat{\beta})}{1+exp(X\hat{\beta})}$$

se.bands.logit = cbind(preds$fit+2*preds$se.fit, preds$fit-2*preds$se.fit)

se.bands = exp(se.bands.logit)/(1+exp(se.bands.logit))

plot(age,I(wage>250),xlim=agelims, type="n",ylim=c(0,.2))

```
points(jitter(age), I((wage>250)/5),cex=.5,pch="|", col="darkgrey")
lines(age.grid,pfit,lwd=2, col="blue")
matlines(age.grid,se.bands,lwd=1,col="blue",lty=3)
```



table(cut(age,4))

```
(17.9,33.5]   (33.5,49]   (49,64.5] (64.5,80.1]
        750        1399         779          72
```

fit=lm(wage~cut(age,4),data=Wage)
coef(summary(fit))

```
                          Estimate Std. Error   t value     Pr(>|t|)
(Intercept)              94.158392   1.476069 63.789970 0.000000e+00
cut(age, 4)(33.5,49]     24.053491   1.829431 13.148074 1.982315e-38
cut(age, 4)(49,64.5]     23.664559   2.067958 11.443444 1.040750e-29
cut(age, 4)(64.5,80.1]    7.640592   4.987424  1.531972 1.256350e-01
```