Statistical Learning and Data mining

Homework 11

M052040003 鍾冠毅

1.a. $\forall x \leq \xi$, $f_1(x)$ has coefficients $a_1 = \beta_0$, $b_1 = \beta_1$, $c_1 = \beta_2$, $d_1 = \beta_3$

1.b. $\forall x > \xi$, $f(x)$ has the form of:

$$\begin{aligned}
&\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)^3 \\
=&\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x^3 - 3x^2\xi + 3x\xi^2 - \xi^3) \\
=&(\beta_0 - \beta_4\xi^3) + (\beta_1 + 3\beta_4\xi^2)x + (\beta_2 - 3\beta_4\xi)x^2 + (\beta_3 + \beta_4)x^3
\end{aligned}$$

Thus, $a_2 = \beta_0 - \beta_4\xi^3$, $b_2 = \beta_1 + 3\beta_4\xi^2$, $c_2 = \beta_2 - 3\beta_4\xi$, $d_2 = \beta_3 + \beta_4$

1.c.

$$\begin{aligned}
&f_1(\xi) = \beta_0 + \beta_1\xi + \beta_2\xi^2 + \beta_3\xi^3 \\
&f_2(\xi) = (\beta_0 - \beta_4\xi^3) + (\beta_1 + 3\beta_4\xi^2)\xi + (\beta_2 - 3\beta_4\xi)\xi^2 + (\beta_3 + \beta_4)\xi^3 \\
=&\beta_0 - \beta_4\xi^3 + \beta_1\xi + 3\beta_4\xi^3 + \beta_2\xi^2 - 3\beta_4\xi^3 + \beta_3\xi^3 + \beta_4\xi^3 \\
=&\beta_0 + \beta_1\xi + \beta_2\xi^2 + 3\beta_4\xi^3 - 3\beta_4\xi^3 + \beta_3\xi^3 + \beta_4\xi^3 - \beta_4\xi^3 \\
=&\beta_0 + \beta_1\xi + \beta_2\xi^2 + \beta_3\xi^3
\end{aligned}$$

1.d.

$$\begin{aligned}
&f'(x) = b_1 + 2c_1 x + 3d_1 x^2 \\
&f_1'(\xi) = \beta_1 + 2\beta_2\xi + 3\beta_3\xi^2 \\
&f_2'(\xi) = \beta_1 + 3\beta_4\xi^2 + 2(\beta_2 - 3\beta_4\xi)\xi + 3(\beta_3 + \beta_4)\xi^2 \\
=&\beta_1 + 3\beta_4\xi^2 + 2\beta_2\xi - 6\beta_4\xi^2 + 3\beta_3\xi^2 + 3\beta_4\xi^2 \\
=&\beta_1 + 2\beta_2\xi + 3\beta_3\xi^2 + 3\beta_4\xi^2 + 3\beta_4\xi^2 - 6\beta_4\xi^2 \\
=&\beta_1 + 2\beta_2\xi + 3\beta_3\xi^2
\end{aligned}$$

1.e.

$$\begin{aligned}
&f''(x) = 2c_1 + 6d_1 x \\
&f_1''(\xi) = 2\beta_2 + 6\beta_3\xi \\
&f_2''(\xi) = 2(\beta_2 - 3\beta_4\xi) + 6(\beta_3 + \beta_4)\xi \\
=&2\beta_2 + 6\beta_3\xi
\end{aligned}$$

2.a. $\hat{g}(x) = 0$, the large smoothing parameter λ forces $g^{(0)} \to 0$

2.b. $\hat{g}(x) = c$, the large smoothing parameter λ forces $g^{(1)} \to 0$

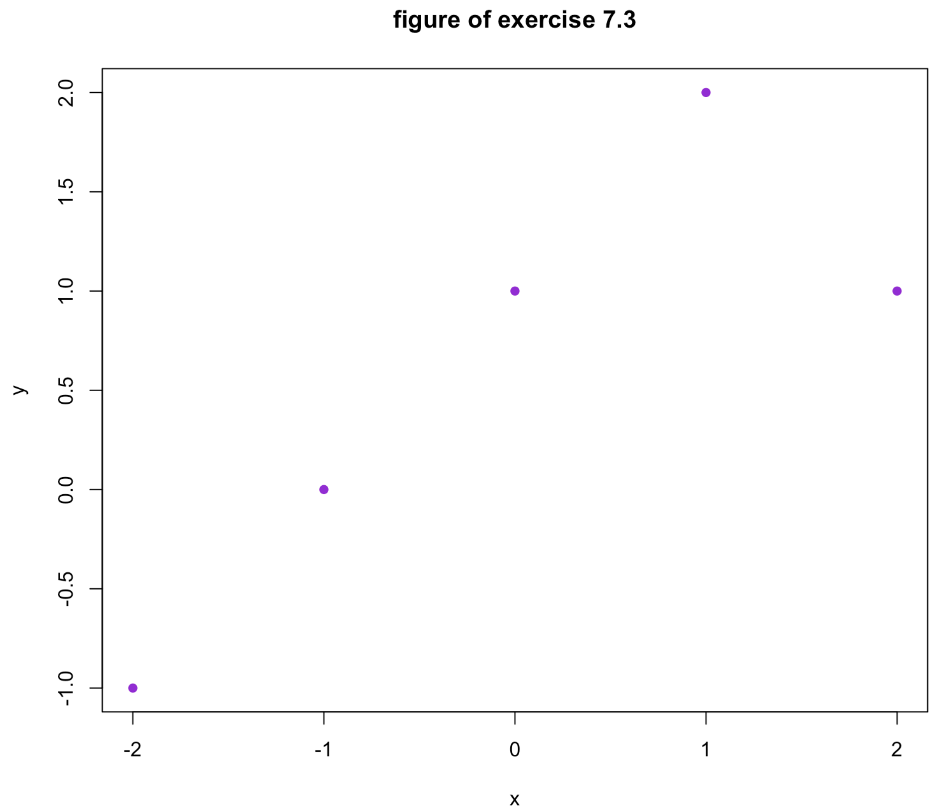
2.c. $\hat{g}(x) = bx + c$, the large smoothing parameter λ forces $g^{(2)} \to 0$

2.d. $\hat{g}(x) = ax^2 + bx + c$, the large smoothing parameter λ forces $g^{(3)} \to 0$

2.e.

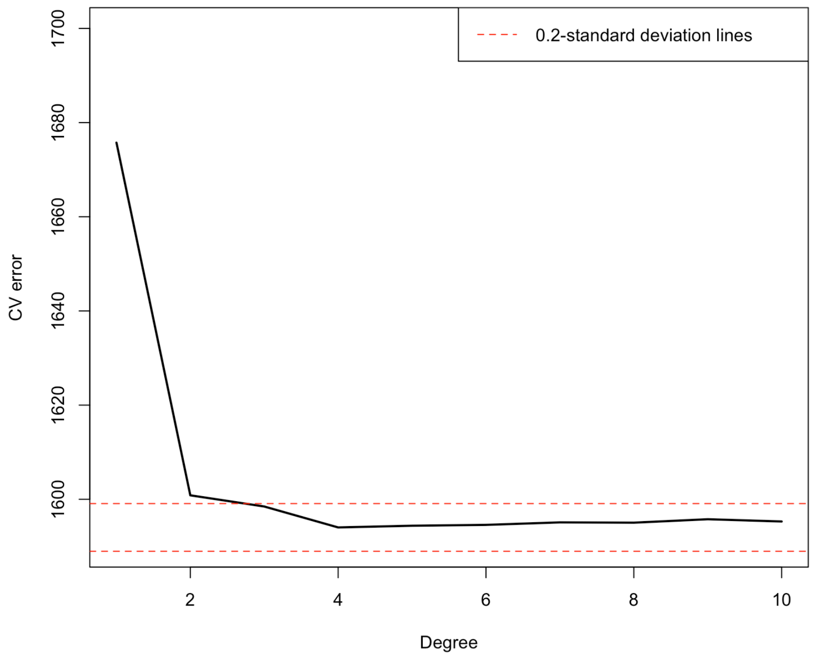
The penalty term no longer matters. This is the formula for linear regression, to choose $\hat{g}$ based on minimizing RSS.

3.



For $x \in [-2, 1)$, $y = 1 + x$ with the slope is 1 and the intercept is 1. For $x \in [1, 2]$, $y = 1 + x - 2(x-2)^2 = -2x^2 + 5x - 1$ which is a quadratic concave curve.

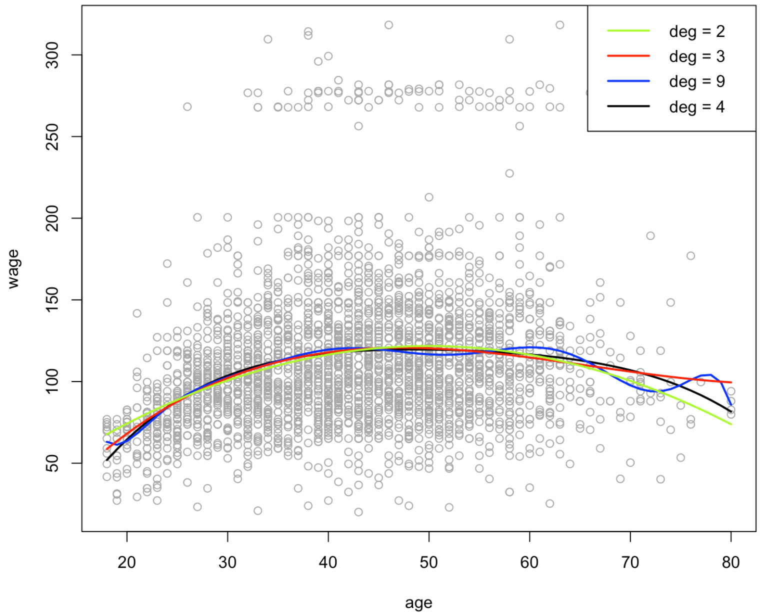
6.a.



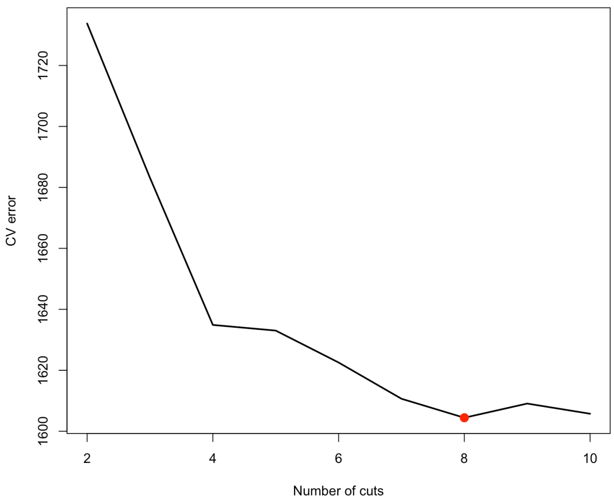
The cv-plot with standard deviation lines show that $d = 4$ is the smallest degree giving reasonably small cross-validation error. We now find best degree using ANOVA.

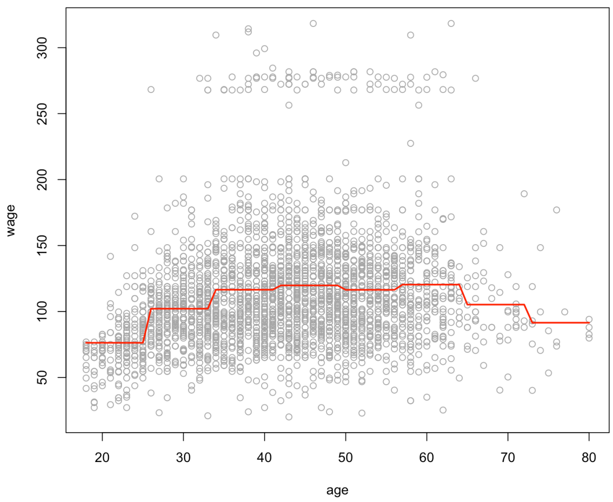
```
> anova(fit.1, fit.2, fit.3, fit.4, fit.5, fit.6, fit.7, fit.8, fit.9, fit.10)
Analysis of Variance Table

Model  1: wage ~ poly(age, 1)
Model  2: wage ~ poly(age, 2)
Model  3: wage ~ poly(age, 3)
Model  4: wage ~ poly(age, 4)
Model  5: wage ~ poly(age, 5)
Model  6: wage ~ poly(age, 6)
Model  7: wage ~ poly(age, 7)
Model  8: wage ~ poly(age, 8)
Model  9: wage ~ poly(age, 9)
Model 10: wage ~ poly(age, 10)
   Res.Df     RSS Df Sum of Sq        F    Pr(>F)    
1    2998 5022216                                    
2    2997 4793430  1    228786 143.7638 < 2.2e-16 ***
3    2996 4777674  1     15756   9.9005  0.001669 ** 
4    2995 4771604  1      6070   3.8143  0.050909 .  
5    2994 4770322  1      1283   0.8059  0.369398    
6    2993 4766389  1      3932   2.4709  0.116074    
7    2992 4763834  1      2555   1.6057  0.205199    
8    2991 4763707  1       127   0.0796  0.777865    
9    2990 4756703  1      7004   4.4014  0.035994 *  
10   2989 4756701  1         3   0.0017  0.967529    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the anova table, the polynomial model with degree $d = 2$ is the most significant. Now we plot the four polynomials as the figure showed below. The four curve almost consist.



6.b.

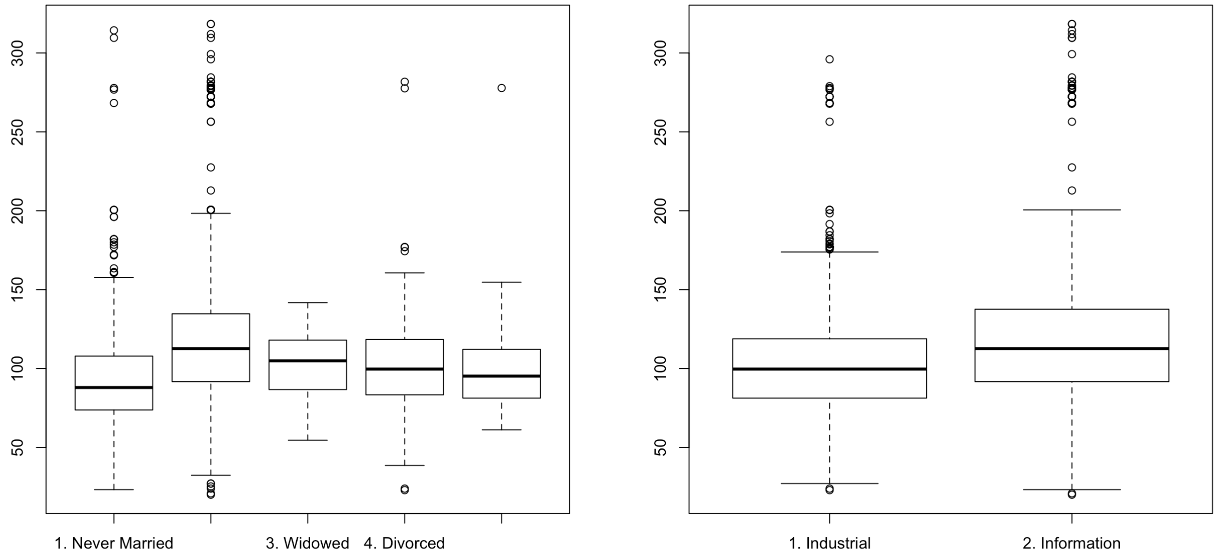




The cross validation shows that test error is minimum for $k = 8$ cuts. We now train the entire data with step function using 8 cuts and plot it.

7.

It appears a married couple makes more money on average than other groups. It also appears that Informational jobs are higher-wage than Industrial jobs on average.




```
> fit <- gam(wage ~ maritl + jobclass + s(age, 4), data = Wage)
> deviance(fit)
[1] 4476501
> fit <- lm(wage ~ maritl, data = Wage)
> deviance(fit)
[1] 4858941
> fit <- lm(wage ~ jobclass, data = Wage)
> deviance(fit)
[1] 4998547
> fit <- lm(wage ~ maritl + jobclass, data = Wage)
> deviance(fit)
[1] 4654752
> fit <- gam(wage ~ maritl + jobclass + s(age, 4), data = Wage)
> deviance(fit)
[1] 4476501
```

The GAM method perform the best for its smallest deviance among the five models.