3.

$(a) a_1 = \beta_0, b_1 = \beta_1, c_1 = \beta_2, d_1 = \beta_3$

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

$(c)(d)(e)$Notice that $f_1(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$, $f_2(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4(x-\xi)^3$

$f_2(x) - f_1(x) = \beta_4(x-\xi)^3$

Apprently $f_1(\xi) = f_2(\xi)$, since the term vanishes.

$f_2'(x) - f_1'(x) = 3\beta_4(x-\xi)^2$

$f_2''(x) - f_1''(x) = 6\beta_4(x-\xi)$

Similarly, $f_1'(\xi) = f_2'(\xi)$ and $f_1''(\xi) = f_2''(\xi)$ hold since the above terms vanish.

4.

$(a)(b)$As $\lambda$ approaches infinity, the second term, the penalty term, of the loss function dominates, and it has to be minimized to achieve the argmin, i.e., the $\hat{g}_{1,2}$. Therefore, the penalty term has to be zero, which means the $\hat{g}_1$ is a polynomial of order 2 (or less), and $\hat{g}_2$ order 3 (or less), that minimize the first term.

Therefore, $\hat{g}_2$ has less training error than $\hat{g}_1$.

Generally speaking, $\hat{g}_1$ should be smoother than $\hat{g}_2$, and thus is more likely to prevent overfitting, and thus $\hat{g}_1$ is likely to have less test RSS.