STAT 745 – Fall 2014
Assignment 3

Group 1: Francene Cicia, Doug Raffle, Melissa Smith

September 15, 2014

1 Loess

1.a Find θ_0 in loess $d = 0$.

$$\min_{\theta_0} \sum_{i=1}^{n} K_\lambda(x_0, x_i)(y_i - \theta_0)^2$$

We start by differentiating with respect to θ_0:

$$\frac{\partial}{\partial \theta_0} \sum_{i=1}^{n} K_\lambda(x_0, x_i)(y_i - \theta_0)^2 = \sum_{i=1}^{n} 2K_\lambda(x_0, x_i)(y_i - \theta_0) = 2\sum_{i=1}^{n} K_\lambda(x_0, x_i)y_i - 2\theta_0 \sum_{i=1}^{n} K_\lambda(x_0, x_i)$$

Then set the derivative equal to zero and solve for θ_0:

$$2\sum_{i=1}^{n} K_\lambda(x_0, x_i)y_i - 2\theta_0 \sum_{i=1}^{n} K_\lambda(x_0, x_i) = 0$$

$$\hat{y} = \hat{\theta_0} = \frac{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i)y_i}{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i)}$$

1.b Find θ_0 and θ_1 in loess $d = 1$

$$\min_{\theta_0, \theta_1} \sum_{i=1}^{n} K_\lambda(x_0, x_i)(y_i - \theta_0 - \theta_1 x_i)^2$$

We can start by finding taking partial derivatives with respect to θ_0 and θ_1, setting them to zero, and solving for their corresponding θ_i.

$$\frac{\partial}{\partial \theta_0} \sum_{i=1}^{n} K_\lambda(x_0, x_i)(y_i - \theta_0 - \theta_1 x_i)^2 = -2\sum_{i=1}^{n} K_\lambda(x_0, x_i)(y_i - \theta_0 - \theta_1 x_i)$$

$$0 = -2\sum_{i=1}^{n} K_\lambda(x_0, x_i)(y_i - \theta_1 x_i) + 2\theta_0 \sum_{i=1}^{n} K_\lambda(x_0, x_i)$$

$$\hat{\theta_0} = \frac{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i)y_i - \theta_1 \sum\limits_{i=1}^{n} K_\lambda(x_0, x_i)x_i}{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i)} \tag{1}$$

$$\frac{\partial}{\partial \theta_1} \sum_{i=1}^{n} K_\lambda(x_0, x_i)(y_i - \theta_0) - \theta_1 x_i)^2 = -2 \sum_{i=1}^{n} x_i K_\lambda(x_0, x_i)(y_i - \theta_0 - \theta_1 x_i)$$

$$0 = -2 \sum_{i=1}^{n} x_i K_\lambda(x_0, x_i)(y_i - \theta_0) + 2\theta_1 \sum_{i=1}^{n} K_\lambda(x_0, x_i) x_i^2$$

$$\hat{\theta_1} = \frac{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i y_i - \theta_0 \sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i}{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i^2} \tag{2}$$

Since $\hat{\theta_0}$ is written in terms of θ_1 and vice versa, we can use $\hat{\theta_1}$ from Equation 2 to solve for $\hat{\theta_0}$ explicitly.

$$\hat{\theta_0} = \frac{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) y_i - \left(\frac{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i y_i - \theta_0 \sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i}{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i^2} \right) \sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i}{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i)}$$

Which simplifies to:

$$\hat{\theta_0} = \frac{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i^2 \sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) y_i - \sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i \sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i y_i}{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) \sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i^2 - \left(\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i \right)^2} \tag{3}$$

This estimator can now be used in Equation 2 for θ_0

$$\hat{\theta_1} = \frac{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i y_i - \hat{\theta_0} \sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i}{\sum\limits_{i=1}^{n} K_\lambda(x_0, x_i) x_i^2} \tag{4}$$

Equations 3 and 4 can be used to find the coefficients in our final estimator:

$$\hat{y} = \hat{\theta_0} + \hat{\theta_1} x$$

2 Least Squares and Ridge Regression

2.a Least Squares

$$H = X(X^T X)^{-1} X^T$$

Estimator:

$$\begin{aligned} \hat{f} &= X\hat{\beta}^{(ls)} \\ &= X(X^T X)^{-1} X^T y \\ &= Hy \end{aligned}$$

Trace:

$$tr(H) = \sum_{i=1}^{n} H_{ii} = rank(X) = p$$

The trace of H is equal to the number of parameters in Least Squares Regression.

2.b Ridge

$$H_\lambda = X(X^T X + \lambda I)^{-1} X^T$$

Estimator:

$$\begin{aligned}\hat{f} &= X\hat{\beta}_\lambda^{(ridge)} \\ &= X(X^T X + \lambda I)^{-1} X^T y \\ &= H_\lambda y\end{aligned}$$

Trace:

$$tr(H_\lambda) = df(\lambda) = \sum_{i=1}^{p} \frac{d_i^2}{d_i^2 + \lambda}$$

Where p is the number of parameters ($rank(X)$), the d_i's are the singular values of X (similarly, the d_i^2's are the eigenvalues of $X^T X$), and $df(\lambda)$ are the effective degrees of freedom.

This result makes intuitive sense in the context of Ridge Regression. Extending principal component analysis, variables with small eigenvalues have less variation and should therefore be less informative about y. Ridge Regression implements this idea by giving variables with small eigenvalues lower weights, according to the ratio $(d_i^2)/(d_i^2 + \lambda)$.

When $\lambda = 0$:

$$H_\lambda = X(X^T X + 0I)^{-1} X^T = X(X^T X)^{-1} X^T = H$$

$$tr(H_\lambda) = \sum_{i=1}^{p} \frac{d_i^2}{d_i^2} = \sum_{i=1}^{p} 1 = p = tr(H)$$

So for $\lambda = 0$, Ridge Regression is equivalent to Least Squares Regression. As λ increases, however, the variables are "shrunk" more and more, and the less informative variables are essentially zeroed out. Doing this removes variables from consideration, which effectively shrinks the parameter space. Since the degrees of freedom for the model depend on the number of parameters, we should adjust them accordingly, hence the *effective* degrees of freedom.