# Homework 9

Nathan Bick

1. Problems 13.F: 1

1. Solve the system of two equations in the two unknowns $R_0$ and $R_1$ in the proof of Lemma 13.3 and show that the solutions are as given in Lemma 13.3.

Lemma 13.3: $\hat{R}_1 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \dfrac{n\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}$

and $\hat{R}_0 = \bar{y} - \hat{R}_1\bar{x}$.

The system of equations is:

$$nR_0 + \left(\sum_{i=1}^{n}x_i\right)R_1 = \sum_{i=1}^{n}y_i$$

$$\left(\sum_{i=1}^{n}x_i\right)R_0 + \left(\sum_{i=1}^{n}x_i^2\right)R_1 = \sum_{i=1}^{n}x_iy_i.$$

first we see

$$\begin{cases} \hat{R}_0 = \sum y_i - (\sum x_i)\hat{R}_1 = \bar{y} - \hat{R}_1\bar{x} \\ (\sum x_i)\hat{R}_0 + (\sum x_i^2)\hat{R}_1 = \sum x_iy_i \end{cases}$$

$$\Rightarrow (\sum x_i)\left(\frac{\sum y_i - (\sum x_i)\hat{R}_1}{n}\right) + (\sum x_i^2)\hat{R}_1 = \sum x_iy_i.$$

$$\Rightarrow \frac{\sum x_i \sum y_i - (\sum x_i)^2\hat{R}_1}{n} + \frac{n}{n}(\sum x_i^2)\hat{R} = \sum x_iy_i$$

$$\Rightarrow \frac{n\left(\sum x_i^2\right) R_1 - \left(\sum x_i\right)^2 R_1}{n} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$\Rightarrow \left(n \sum x_i^2 - \left(\sum x_i\right)^2\right) R_1 = n \sum x_i y_i - \sum x_i \sum y_i$$

$$\Rightarrow R_1 = \frac{n \sum x_i y_i - \sum x_i y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

2. Exercises 14B.1

1. Compare Cases A, B, and C of example 14.2. Which Case is best? Why?

In example 14.2, we see three cases of one-stage clustering. In all three cases we get the same $E(\bar{y}_{clus}) = \frac{9}{2} = \bar{y}_U$, so each case shows an unbiased and effective estimator. However, each case has very different $V(\bar{y}_{clus})$.

A: $V(\bar{y}_{clus}) = 9$, B: $V(\bar{y}_{clus}) = \frac{1}{4}$, C: $V(\bar{y}_{clus}) = 0$.

C is better than B, which is better than A. We see if the clusters are each internally heterogeneous among the y values, $V(\bar{y}_{clus})$ will be less than when each cluster is homogeneous.

C is best because our sampling variance is low.