2.(a). Data Description:

The data are on the homicide rate in Detroit for the years 1961-1973

FTP - Full-time police per 100,000 population

UEMP - % unemployed in the population

MAN - number of manufacturing workers in thousands.

LIC - number of handgun licenses per 100,000 population.

GR - number of handgun registrations per 100,000 population.

~~CLEAR - % homicides cleared~~

NMAN - number of non-manufactoring workers in thousands

GOV - number of government workers in thousands.

HE - Average hourly earnings

WE - Average weekly earnings

HOM - number of homicides per 100,000 of population.

From the problem. we know that we use three of the variables can predict HOM:

Using the basic linear regression function model: $y(x,w) = w_0 + w_1 x_1 + \dots + w_D x_D$. $X = (x_1, \dots x_D)^T$

s.t. $y(X,w) = w_0 + w_1 X_1 + w_2 X_2 + w_3 X_3$. $X_1 = (FTP(1), FTP(2) \dots FTP(13))^T$

$= wX$ $X_2 = (WE(1), WE(2) \dots WE(13))^T$

If we want to find the best third variable: and we need to find X_3.

We can try all the other variables and find the minimize sum-of-squares error $=0$

$E_D(w) = \frac{1}{2}\sum_{n=1}^{N} \{t_n - w^T \phi(x_n)\}^2$ $w = (\Phi^T \Phi)^{-1} \Phi^T t$ Φ is X t is $(HOM(1), HOM(2) \dots HOM(13))^T$

then for different third variable we will get different w.

And we can use RMSE, by finding the minimize RMSE. we can find the right third variable:

$$RMSE = \cancel{\sqrt{\frac{\sum_{i=1}^{N}(HOM(i) - y_i(x,w))^2}{N}}} = \sqrt{\frac{\sum_{i=1}^{n}(y(x(i),w) - HOM(i))^2}{n}}$$

I wrote 2 piece of code one is just use the original data. the other is normalized data by divide its ~~max value~~ upper bound in column. s.t. every value will $\in [0,1]$.

Both of them show that LIC. is the best third variable to choose.

without normalization: $y = -58.1244 + 0.1847\,FTP + 0.1068\,WE + 0.0165\,LIC$. → result in LinearR.m

with normalization: $y = -0.5812 + 0.7388\,FTP + 0.3205\,WE + 0.1482\,LIC$. → result in NLinearR.m

2(b). i: using b to replace ~~figure~~ feature 1's '?'

for real-value just replace missing values with the label-conditioned mean.

for letter-value just replace missilf values with mode value.

St. replace all '?' in feature 2, 3, 15 with mean others with mod. value