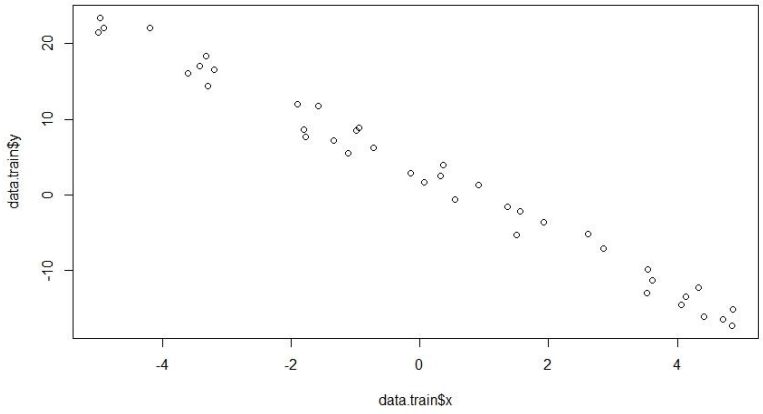
Assignment 4 - Report

1. **Part-1:**
    a. Learning rate alpha is typically between 10 and 0.1. However, in our case, the alpha needs to be even smaller because the comparative large values of variable in data. Keeping a bigger alpha i.e. 0.1 makes the parameters w0, w1 overshoot and become too large. So, I took 0.001 as my alpha value which gave me a mean square error (MSE) of 2.77.
    b. Convergence criteria: ||**oldw-w**|| <0.0001, where **oldw** is (oldw0, oldw1) and **w** is (w0,w1). The idea behind keeping 0.0001 as the threshold is that the value of w has stabilized and won't be changing much in upcoming iterations. I have verified for even smaller convergence rate and as expected, it showed the same results. I even checked the linearity assumption as shown below:



2. **Part-2:**
    a. Learning rate for this case is 0.01 and convergence criteria as 0.00001. The idea behind keeping an even smaller convergence criteria is because the values are normalized and will converge gradually. Similarly, because of **normalization** we don't have to reduce the value of alpha below 0.1 which serves our purpose. Since the **w** is initialized to **0**, the **w** in first iteration will be very small and I had to put up an additional criterion that the minimum iterations should be 1000. Scatter plot didn't prove to be useful in this case as the variables are ordinal in nature.
    b. Final feature set decided is (age, mother's education, father's education, travel time to school, # of past class failure, going out with friends)
    c. I selected the above-mentioned features based on correlation coefficient values. I had selected all those features which had a correlation coefficient of 0.1 or more with the target variable. I had tried to perform feature selection based on the mutual information as well since our variables are of categorical/ordinal nature. The features that were significant in this case are (mother's education, father's education, # of past class failure, travel time to school, going out with friends) which is nearly the same as in case of correlation.

d. I had performed 5-fold cross validation to come up with mean square error. My model gives a MSE value of 0.157 for features selected using correlation coefficient and a MSE of 0.37 for that with mutual information.

3. References:

a. http://asv.informatik.uni-leipzig.de/uploads/document/file.../TMI04.2_linear_regression.pdf