Linear regression, correlation coefficient and significance of regressors.

The datafile "`lead_mortality.csv`" contains data on 172 U.S. cities in 1900. These data were provided by Professor Karen Clay of Carnegie Mellon University and are a subset of the data used in her paper with Werner Troesken and Michael Haines Lead and Mortality, Review of Economics and Statistics, 2014.

```
Download the file ‘‘lead_mortality.csv’’ and open it with excel.
Either transform it to an "xlsx" file and import the data in matlab,
either directly import the data from the csv file to matlab.
```

Lead is toxic , particularly for young children and for this reason government regulations severely restrict the amount of lead in our environment. But this was not always the case. In the early part of the 20th century, the underground water pipes in many U.S. cities contained lead, and lead from these pipes leached into drinking water.

You will investigate the effect of these lead water pipes on infant mortality. The data file "`lead_mortality.csv`", contains data on infant mortality, type of water pipes (lead or non-lead), water acidity (pH) , and several demographic variables for 172 U.S. cities in 1900. Using this data:

a. Compute the average infant mortality rate (Infrate) for cities with lead pipes and for cities with non-lead pipes. Test the normality of each character separately. If the conclusions are good, compare the equality in means of these two normal samples. Previously, test homoscedaticity. If it rejects, build a p-value for a student test of the mean of the first equals to that of the second. Compute the R-squared associated to the regression of Infrate on lead.

b. The amount of lead leached from lead pipes depends on the chemistry of the water running through the pipes. The more acidic the water (that is, the lower its pH), the more lead is leached. Regress Infrate on regressors Lead, pH, and the interaction term Lead×pH. Construct confidence intervals for the regression coefficients. We assume a priori homoscedadicity of the residues in each regressor, as independence. Test the significance of regressors and for those that remain compute the R-squared of the corresponding regression.

c. Interpretation.

i. The regression includes four coefficients (the intercept and the three coefficients multiplying the regressors). Explain what each coefficient measures.
ii. Plot the estimated regression function relating Infrate to pH for Lead = 0 and for Lead =1. Describe the differences in the regression functions and relate these differences to the coefficients you discussed in (i).
iii. Does Lead have a statistically significant effect on infant mortality? Explain.
iv. Does the effect of Lead on infant mortality depend on pH? Is this dependence statistically significant?

Bliographical links (except your online course):
https://eric.univ-lyon2.fr/ ricco/cours/cours/Regression_Lineaire_Multiple.pdf
http://www.math.univ-toulouse.fr/ besse/Wikistat/pdf/st-l-inf-intRegmult.pdf