Homework assignment 1.

Consider data in *HousingData.csv* (attached file). The Boston Housing Dataset is derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA. The following describes the dataset columns:

CRIM:	per capita crime rate by town
ZN:	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS:	proportion of non-retail business acres per town.
CHAS:	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX :	nitric oxides concentration (parts per 10 million)
RM:	average number of rooms per dwelling
AGE:	proportion of owner-occupied units built prior to 1940
DIS:	weighted distances to five Boston employment centres
RAD:	index of accessibility to radial highways
TAX:	full-value property-tax rate per $10,000
PTRATIO :	pupil-teacher ratio by town
B :	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT:	lower status of the population
MEDV:	Median value of owner-occupied homes in $1000's

1. (a) **(0.25 pt.)** Look closely at the database. Some data are missing. What would you recommend to do with this? Explain. Fill in these values.

   (b) **(0.25 pt.)** Construct linear regression model for $MEDV$ taking other variables as explanatory. Test the hypothesis about significance of the variable NOX. Write down the corresponding hypotheses and the test statistics which should be used. Explain how the test works. Write down the critical region for significance level 5%. Run the test and formulate the conclusion.

   (c) **(0.5 pt.)** Select variables using AIC. Explain how the procedure works. For the final model present the output. Interpret the resulting estimates for coefficients. Explain what the F-statistics for the regression shows. How to interpret it?

   (d) **(0.25 pt.)** Test whether the effect of the independent variable $PTRATIO$ on $MEDV$ is negative c.p.. State your intuitive expectation. Formulate the hypotheses. Test them. Interpret the result.

   (e) **(0.25 pt.)** Interpret the coefficient of variable $CHAS$. Explain why we do not include into the model variable NCR =1 if tract does not bound river; 0 otherwise.

   (f) **(0.5 pt.)** Test variables DIS and RAD for **joint** significance. What test to use? Write down the hypotheses, test statistics and critical region for significance level 5%. Run the test. Formulate conclusions.

   (g) **(0.5 pt.)** Test whether the effect of the $CRIM$ on the expected value of $MEDV$ is the same for tracts near Charles River and others. Run the corresponding model. How to interpret the result?

2. (a) **(0.25 pt.)** Construct log-linear model for $log(MEDV)$ taking other variables as independent variables. Select variables using AIC.

   (b) **(0.5 pt.)** Interpret the estimated values of coefficients. Do they correspond to your intuition?

   (c) **(0.5 pt.)** Test whether the expected percentage decrease of $MEDV$ caused by the one point increase of $PTRATIO$ is greater than that caused by one point increase of $CRIM$ c.p.. Write down the hypotheses. Be careful with the sign of interest. Explain what test to use and how it works. Run the test. Formulate the conclusions.

   (d) **(0.25 pt.)** Can we compare linear and log-linear model using adjusted R^2 coefficients? Why? Explain.