Regression analysis of *wages* dataset

Statistical Learning - First Module

Description of the dataset

The dataset *wages* is a subset of a survey for the U.S. which provides data on the workers' looks as well as on labor-market and demographic variables. The variables are

- *wage*: hourly wage
- *exper*: years of workforce experience
- *looks*: ranking made by an interviewer for physical attractiveness, using five categories (homely, quite plain, average, good looking, and strikingly beautiful or handsome) coded from 1 to 5, respectively.
- *union*: if union member (yes/no)
- *goodhlth*: if good health (yes/no)
- *educ*: years of schooling
- *ethnicity*: ethnicity (black/white)
- *gender*: gender (male/female)
- *marital*: marital status
- *region*: if the person lives in a northern or southern state
- *city*: if the person lives in a small, medium or big city
- *industry*: if the person works in service or manufacturing industry

Aim of the regression analysis

Use the linear regression model, with *wage* as the dependent variable, to estimate the relationship between employees' earnings and their socio-demographic characteristics. Can you find evidence of some form of discrimination patterns? In particular,

- find the best model specification (suggestions: try useful transformations or rescalings of the variables and consider that *exper* and *edu* may exert a non-linear effect and that *looks* may be re-categorised).

- report the proper interpretation of the model parameters
- perform an inferential analysis through the proper hypothesis test procedures and confidence intervals. Is there convincing evidence that women with above average looks earn more than women with average looks? Does physical appearance exert the same effect on the wage for male and female employees? Is the return to education, in terms of wage, the same for black and white workers?
- verify whether there are important violations of the model assumptions
- if violations of the homoscedasticity assumption are detected, perform the proper remedials. Can the FWLS approach be useful in dealing with the problem?