Linear models for Gaussian data

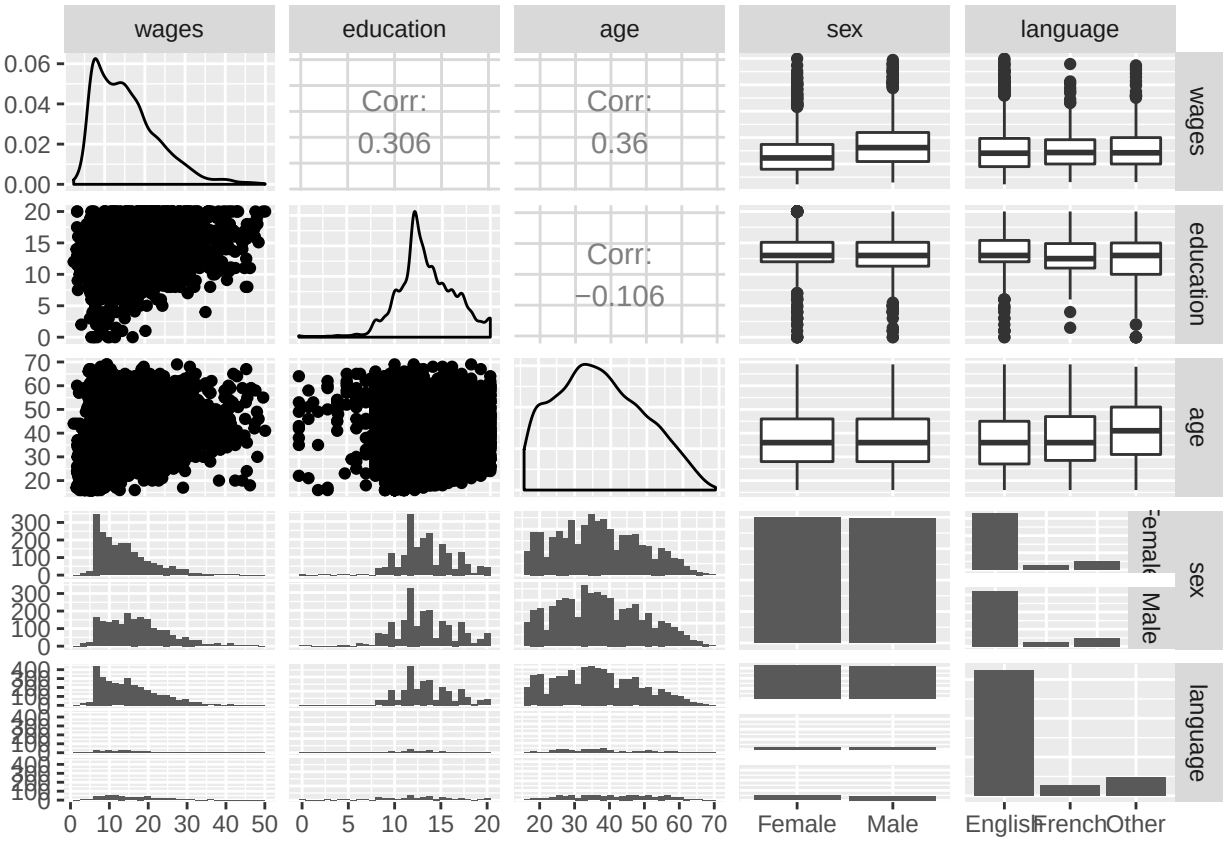
TMA4315 - Exercise 1

Anders Christiansen Sørby, Edvard Hove

We will work on the dataset `carData`, which consists of 3987 observations on the following 5 variables:

- `wages`, composite hourly wage rate from all jobs
- `education`, number of years in schooling
- `age`, in years
- `sex`, Male or Female
- `language`, English, French or Other

Using the function `ggpairs` we get the following diagnostic plot matrix:



The plot suggests that `wages` increase with both `education` and `age`, despite the apparent slight decrease of `education` with `age`. It also seems like males recieve higher wages than females, despite similar levels of age and education. There seems to be some relationship between `language` and `age`, but it is not clear if there is a corresponding relationship between `language` and `wages`.

When performing a multiple linear regression analysis to study how `wages` depends on the explanatory variables we need to assume that: ELABORATE

- each data point Y_i is independent
- there is a linear relationship between `wages` and the explanatory variables, such that $\mu_i = \eta_i = \mathbf{x_i}^T\beta$)
- the residuals are normally distributed with a homogenous variance, such that $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$