Prestige data Assignment

We have a data set containing information about 102 professions; load the libraries `car` and `carData` and type `?Prestige` to see a detailed description of the data. Define `data=na.omit(Prestige)`, to make sure that there are no missing values in your data set. We will consider `prestige` to be the dependent variable in our model, and `education, income` and `type` as the covariates (we will not use the variables `women` and `census`).

(a) Fit the model containing only the main effects, and in this model, test the hypothesis that the effect on prestige of an occupation being White Collar, is the same as the effect of the occupation being Professional, at a significance of 5%.

(b) In the same model, test the hypothesis that `type` has no effect on the prestige of an occupation. again at the 5% significance level. Also, test the hypothesis that the regression constant for `income` is given by 0.002 and the regression constant for `education` is given by 2.

(c) Now consider the model where `income` and `education` are allowed to have interactions with `type`. Explain whether or not you would prefer this model over the model used in (a) and (b).

(d) In the model introduced in (c), test whether `income` has no effect on prestige for white collar occupations. Hint: carefully consider how income plays a role for white collar occupations in the model with interactions!

(e) Predict the prestige of a new occupation, that has an average income of 5000$, average education of 12 and is a white collar occupation, based on the model in (c).

(f) Test in the model in (c) whether the interactions between `education` and `type` are significant, at a 5% significance level.

We have seen that there are many models possible. Suppose we consider the model in (c) as the largest possible model (i.e., the model with the most free parameters, or highest p). What is the value of p in this case? We can ask ourselves which subset of these p covariates would constitute the 'best" model.

Alice suggest the following method: we start with the full model, see which of the covariates has the highest p-value, and if this value is higher than 5%, we leave it out of the model and repeat these steps for the model with $p-1$ covariates, until all covariates are significant.

Bob, however, suggests a different method: he starts by fitting p models, each with only one covariate, and checks which covariate is the most significant. Then he takes this covariate, and fits all the $p-1$ models with one extra covariate, and he picks the second covariate by checking which one is the most significant. He repeats this process until each added covariate would not be significant any more (so all non-selected covariates would have a p-value greater than 5% when added to the current model).

(g) We would like to compare the two methods. Start by fixing a model without interactions, so find estimates for all parameters in this model (including the variance of the noise!); this will be the model you will simulate from, using the given values of the covariates, but getting new realisations of the dependent variable `Prestige`. Then perform both Alice's and Bob's method to find the "best" model and determine the predicted prestige by these two models for the new occupation considered in (e). Compare these two predictions to the "truth", which is the prediction given by the model you simulate from. Finally, repeat these steps a lot of time and give your conclusions.

This is a challenging exercise! Try to motivate your conclusions by using insightful pictures of your simulation results. Also, it might be helpful to use the command `model.matrix` from the `stats`-package in the full model, as this would give you an easy way to leave out or add covariates (since they correspond to columns of the X-matrix).