Assignment 5: Multiple linear regression; regression with groups

Due at 12 noon, 22 September

Download the `prestige.csv` data set from Learn and create an RStudio project for the assignment. Read the data and keep a sample of 95 observations based on your student ID number.

The `prestige.csv` file contains data for a several-decades old Canadian survey of 98 occupations, including variables like `prestige` (a measure of social standing, our response), `education` (years of formal education), `income` (annual income in dollars), and job_type (classified as tradie, white collar and professional).

1. Take a sample of 95 observations based on your student ID and call it `my_jobs`, adjust the `income` variable to account for inflation (multiplying the values by 10), convert the `job_type` to a factor (using `factor(job_type)` in mutate), and create a variable `log_income`, which is the `log10()` of income).

2. Create a scatterplot of `prestige` versus `education` and include the simple linear regression trend in the plot. Then fit a linear regression model for `prestige` on `education` (call it `m1`). Discuss the residuals for `m1`.

3. Create a scatterplot of `prestige` versus `income`, and another one with `prestige` versus `log_income` include the simple linear regression trend in the plot. Then fit a linear regression model for `prestige` on `log_income` (call it `m2`). Discuss the residuals for `m2`. Why do you think log_income would perform slightly better than income?

4. We are now interested in checking if the relationship between the variables is constant across `job_type`. Create another scatterplot of `prestige` versus `education`, with a different colour for each `job_type`. Before fitting other models, how likely you think it is that the regression lines differ between seasons? (15 words)

5. Fit regressions of `prestige` versus `education` for each of the `job_types`, with common slope but different intercepts (`m3`). Explain in no more than 70 words the change of fit from m1 through to m3. Compare the residual plots for `m1` and `m3`.

6. Write down the regression line coefficients for each of the `job_types` in `m3`.

7. Expand model `m3` to also include `log_income` and call it `m4`. Does the model fit improve from `m3` to `m4`? Do the residuals seem to better fit the assumptions?

8. Include all the answers and code used to obtain them in a Word document and submit it through Learn. Remember to use your own words when answering the questions.