FIN 373 Homework 8

due: **4/3/24**

Instructions: Please submit solutions on canvas. Only a knitted pdf of an `Rmarkdown` file will be accepted.

Problem 1: In this exercise we're going to look at the effect of a educational television program The Electric Company that ran from 1971-1977 on children's reading scores. We will investigate what reading gains, if any, were made by the 1st through 4th grade classes as part of a randomized experiment.[1]

The data comes from a two location trial in which treatment was randomized at the level of school classes.[2] Each class was either treated (to watch the program) or control (to not watch the program). The outcome of interest is the score on a reading test administered at the end of each year called `post.score`. Note that these are distinct classes from all four years. The variables in `electric-company.csv` are:

Variable	*Description*
`pair`	The index of the treated and control pair (ignored here).
`city`	The city: Fresno (`"F"`) or Youngstown (`"Y"`)
`grade`	Grade (1 through 4)
`supp`	Whether the program replaced ("R") or supplemented ("S") a reading activity
`treatment`	"T" if the class was treated, "C" otherwise (randomized)
`pre.score`	Class reading score *before* treatment, at the beginning of the school year
`post.score`	Class reading score at the end of the school year

a. Read the data into an data frame named `electric`. What sort of variable has R assumed `grade` is? How will it be treated in a linear model if we use it as an independent variable? Under what circumstances would that be reasonable or unreasonable?

   Make a new variable from `grade` that is a factor. How will a linear model treat this new variable? (*Hint*: You may find that `summary()` illuminates the new data set).

   Finally, overwrite the existing `treatment` variable so that it is numerical: 1 when the class is treated and 0 when not.

b. Let's now consider the effect of `treatment`. First, fit a linear model that predicts `post.score` with just the `treatment`. Then fit a model uses your factor version of `grade` as well as `treatment`.

   Summarize both models in terms of how much of the variance in `post.score` they "explain".

   Now, consider each model's `treatment` coefficient. Are the estimates of this coefficient *different* in the two models? Why do you think that is?

c. Now make another model that uses the factor version of `grade` and `pre.score` (the reading score before the year begins) to predict `post.score`. Is this model better? If so, in what ways?

[1] This exercise is based on The Electric Company: Television and Reading, 1971-1980: A Mid-Experiment Appraisal. Joan G. Cooney (1976). Children's Television Network Report.

[2] Classes were paired, but we will ignore that in the analysis.

d. Now let's consider the effect of treatment *within* each grade. We can use the `lm` function's "subset" argument to fit the model on just a subset of all the rows in the data set. For example, we can fit a model of the relationship of `post.score` to `treatment` and `pre.score` just in grade 2 like this:

```
mod <- lm(post.score   treatment + pre.score, data = electric, subset = grade == 2)
```

Fit a linear model predicting `post.score` using `treatment` and `pre.score` for each grade. Follow the procedure below:

(a) Define a function named `fit_reg()` that returns the coefficient on `treatment`. The function should have two arguments: the entire data (`data_all`) and the grade (`grade_subset`).

(b) Use a for loop and call the `fit_reg()` function for each grade (1 to 4).

(c) Print out the coefficient on treatment using the `print()` function. Store what the `fit_reg()` function returns in a variable.

(d) Briefly comment on the result. There are now *four* treatment effects. How do they differ as grade increases?

e. Finally, let's investigate the separate grade effects in a single model. One way to do this is to *interact* `treatment` with `grade`. Here's a general modeling principle: If you think the *effect* of variable `A` varies according to the *values* of variable `B`, then you should consider adding an interaction between `A` and `B` in your model. In the `lm()` function this amounts to adding an `A:B` term. For example, if `A` and `B` interact to predict Y, then the formula would be, `Y ~ A + B + A:B`, which would fit the model

$$Y_i = \beta_0 + \beta_A A_i + \beta_B B_i + \beta_{AB}(A_i \times B_i) + \epsilon_i.$$

An alternative syntax to fit this model is to use the right-hand-side expression `A*B`. So, to fit the model above using this notation the formula is `Y ~ A * B.` This will automatically include the `A`, `B`, and `A:B` terms in the model!

Fit a model of all the grades that includes `pre.score`, `treatment`, `grade` (factor version), the factor version of `grade` interacted with `treatment`, and the factor version of `grade` interacted with `pre.score` (this is called a fully interacted model). How would you construct grade-specific treatment effects from these coefficients? Show an example for grade 2.