Adding Interaction terms
Multiple linear regression – Stat 230

In this class, you will extend the multiple linear regression model to include allow for the slope of one variable to depend on the value of another variable. You will consider how to do this, how to interpret the results, and how to determine whether the addition of an interaction term discernibly improves the model (i.e., explains discernibly more variability in the response).

In this activity we will begin by working with the `4_8Cyl` data set described in Sections 3.10. You can load the data with the following code

```
cars <- read.csv("https://aloy.rbind.io/kuiper_data/4_8Cyl.csv")
```

Note that there are only two values of `Cyl` (4 and 8).

Part 1. Thinking geometrically

Task 1. Based on the below equation, write down (by hand is fine) the regression equation for 4-cylinder engines and for 8-cylinder engines. Is this an example of the parallel lines, different slopes, or separate lines model?

$$\mu(\texttt{Price}|\texttt{Mileage}, \texttt{Cyl}) = \beta_0 + \beta_1\texttt{Mileage} + \beta_2\texttt{Cyl}$$

Task 2. Based on the below equation, write down (by hand is fine) the regression equation for 4-cylinder engines and for 8-cylinder engines. Is this an example of the parallel lines, different slopes, or separate lines model?

$$\mu(\texttt{Price}|\texttt{Mileage}, \texttt{Cyl}) = \beta_0 + \beta_1\texttt{Mileage} + \beta_2\texttt{Cyl} + \beta_3\texttt{Mileage} \times \texttt{Cyl}$$

Part 2. Fitting an interaction model

> **i** Interaction terms in R
>
> To add interaction terms in R you multiply the terms together in the model formula. For example, the above model formula would is `Price ~ Mileage + Cyl + Mileage:Cyl`.
> Additionally, R follows the conventional modeling advice to always include the main effects for all variables in an interaction term. Thus, `Price ~ Mileage*Cyl` implements the same model ad the previous formula.

Task 3. Fit the following multiple linear regression model and report the fitted regression equation.

$\mu(\texttt{Price}|\texttt{Mileage}, \texttt{Cyl}) = \beta_0 + \beta_1\texttt{Mileage} + \beta_2\texttt{Cyl} + \beta_3\texttt{Mileage} \times \texttt{Cyl}$

Task 4. Assuming `Mileage` $= 10000$, explain how increasing from four to eight cylinders will impact the estimated price.

> **i** Plotting an interaction model
>
> Last class we used the `ggpredict()` and `plot()` command from the `{ggeffects}` R package to plot the parallel lines model. The syntax for `ggpredict()` is shown below. Remember that you need to pass in a fitted mode (`fm`) and specify what terms you want represented in the plot. Here, you need to select the term you will plot on the x-axis as well as any terms involved in adjusting the slope and intercept (i.e., include categorical variables and those involved in interactions).
>
> ```
> fm_preds <- ggpredict(fm, terms = c("x1", "x2"))
> ```
>
> Then, you can just run `plot(fm_preds)` to visually display the model. If you want to display the points as well as the fitted regression lines, add the argument `show_data = TRUE` to the `plot()` command.
> **Setting values to display**. Sometimes `ggpredict()` will select too many levels for the plot (i.e., will display too many lines) to be interpretable, or will choose unimportant values of `x2` to plot. To specify the specific values of the `x2` variable use to form the lines, you can adjust the syntax to
>
> ```
> fm_preds <- ggpredict(fm, terms = c("x1", "x2[a, b, c]"))
> ```
>
> Where the values inside the `[]` specify the specific values of `x2` to use.

Task 5. Use the `ggpredict()` and `plot()` functions to plot the fitted interaction model.

```
fm_preds <- ggpredict(fm, terms = c("Mileage", "Cyl"))
plot(fm_preds, show_data = TRUE)
```

Task 6. Write the fitted model equation for each `Cyl` group. In other words, write down equations for the lines you just plotted in the previous task.

Do we need the interaction?

Task 7. Conduct an appropriate hypothesis test to determine if the interaction term discernibly improves the model. (You can report the hypotheses, test statistic, degrees of freedom, and p-value by hand.)