MATHS 7107
Data Taming
Practical 8

Separate, parallel and identical lines models

1 Preliminaries

- Set up a project in `RStudio`
- Download the `movies.xlsx` file to a `data` subdirectory of your project directory
- Now load the packages
  - `tidyverse`
  - `readxl`
  - `car`
  - `modelr`
- Read in the movie data.

1.1 Aim of today's prac

We are going to build separate, parallel and identical lines models. It turns out that the identical lines models are exactly the same as just fitting a single quantitative variable in a simple linear regression. We've already been doing that for a while, so we might as well start there.

2 Revision of linear models

We'll start by just fitting a simple linear model of as we've been doing for a few weeks. The model with be `score` against `runtime`. (This will turn out to be the **identical lines** model.)

Questions:

1. Graphically represent the relationship between `score` and `runtime`? *(Hint: from Week 6.)*
2. Fit a linear model to the data with `score` as the response variable, and `runtime` as the predictor? Name it `M1`.
   - Use `lm(score ∼ runtime)`
3. Write down the linear model as an equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

   making sure you define the variables y_i and x_i.

3 A categorical model

Now let's look at the relationship between `score` and `genre`.

Questions:

4. What type of variable is `genre`? Convert the variable to the correct type. (We will need a `<fct>` data type for the `lm()` command to work properly, so we might as well convert it now.)
5. What sort of plot should we use to compare `score` and `genre`? Build one in `R`.
6. Fit a linear model between `score` and `genre`.
    - Use `lm(score ~ genre, data = movies)`.
7. Use `model_matrix(movies, ~genre)` to identify the reference level. Which one is it?
8. Use the model summary to write the linear model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 c_{1,i} + \hat{\beta}_2 c_{2,i} + \hat{\beta}_3 c_{3,i}$$

    Make sure you define the variables $c_{1,i}, c_{2,i}, c_{3,i}$. *(We've used the pronumeral c to indicate that these are **categorical**.)*

4 Parallel lines model

Now we'll combine the categorical variable with the quantitative one (with no interactions), and this will give us a **parallel lines** model.

Questions:

9. Graphically represent the relationship between score, run time AND genre? *(Hint: use colour for genre.)*
10. Fit a parallel lines model? Name it `M2`. *(Hint: no interactions.)*
    - Use `lm(score ~ runtime + genre)`
    - Use `model_matrix(movies, ~ runtime + genre)` to see if the reference level is still the same.
11. Use `summary(M2)` to write the model coefficients:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 c_{1,i} + \hat{\beta}_3 c_{2,i} + \hat{\beta}_4 c_{3,i}$$

12. For each of the four levels in the `genre` variable, write down the corresponding line.

$$\begin{aligned} \hat{y}_{action,i} &= \cdots \\ \hat{y}_{animation,i} &= \cdots \\ \hat{y}_{biography,i} &= \cdots \\ \hat{y}_{comedy,i} &= \cdots \end{aligned}$$

    These are the **parallel lines**.
13. Are the lines in Q12 actually parallel? (Do they have the same slope?)
14. Use `Anova(M2)` to see if both predictors are significant.

5 Separate lines model

Finally, we'll combine the categorical variable with the quantitative one, and include interactions between them. This will give us a **separate lines** model.

Questions:

15. Now fit a separate lines model? Name it `M3`. *(Hint: include interactions.)*
    - Use `lm(score ~ runtime + genre + runtime:genre)`
16. Use `summary(M3)` to write the model coefficients:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 c_{1,i} + \hat{\beta}_3 c_{2,i} + \hat{\beta}_4 c_{3,i} + \hat{\beta}_5 x_i c_{1,i} + \hat{\beta}_6 x_i c_{2,i} + \hat{\beta}_7 x_i c_{3,i}$$

17. For each of the four levels in the `genre` variable, write down the corresponding line.

$$\begin{aligned} \hat{y}_{action,i} &= \cdots \\ \hat{y}_{animation,i} &= \cdots \\ \hat{y}_{biography,i} &= \cdots \\ \hat{y}_{comedy,i} &= \cdots \end{aligned}$$

These are the **separate lines**. Are they indeed non-parallel?

6 Evaluating and using the models

Questions:

18. Which model should be choose? Use `Anova()`.
19. Check the assumptions for the best model.
20. Using the best model predict the score for a 2 hour comedy movie. (With 99% confidence and prediction intervals.)