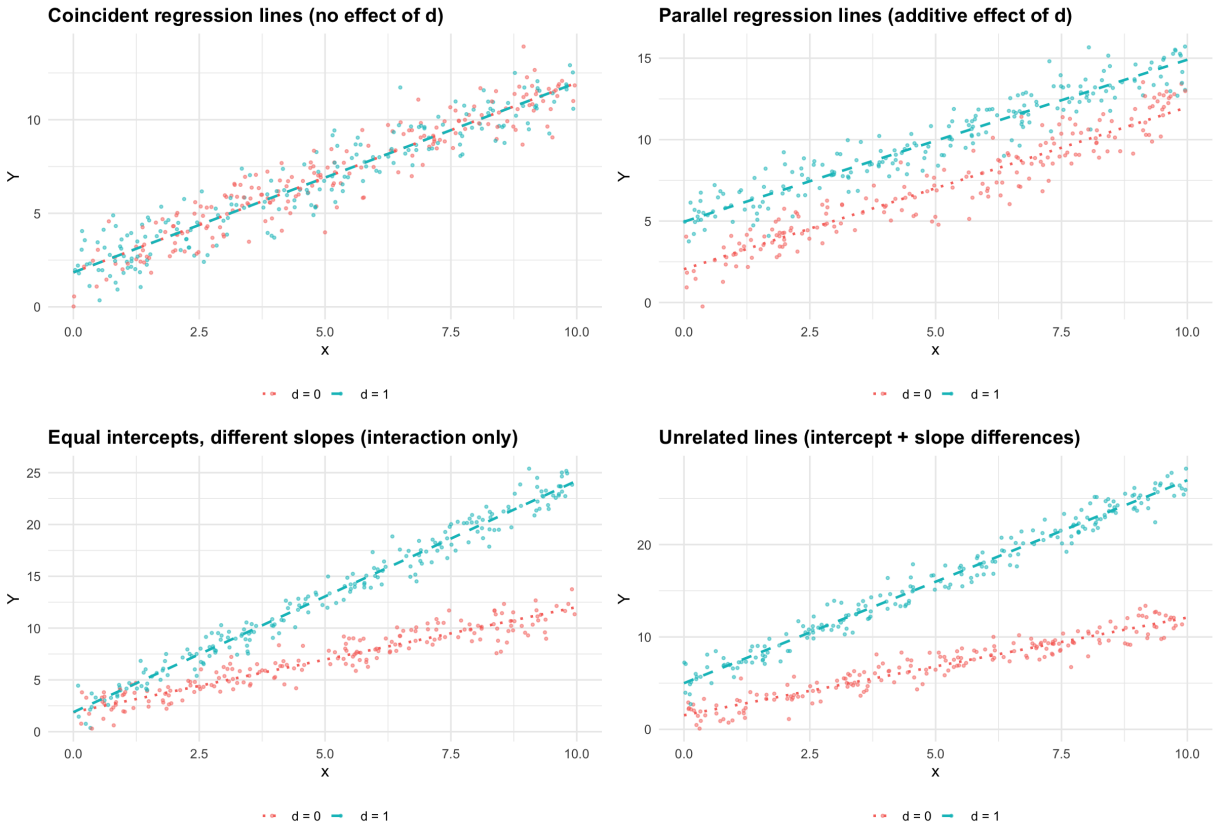
Consider the situation in which we want to model a response variable Y based on a continuous predictor x and a dummy variable d. Suppose that the effect of x on Y is linear. This situation is the simplest version of what is commonly referred to as Analysis of Covariance (ANCOVA), since the predictors include both quantitative variables and qualitative variables.

Forms of regression lines with a dummy variable

– When a regression model includes both a continuous predictor and a dummy variable, the effect of the dummy can take several forms. It may have no impact (coincident lines), shift the intercept (parallel lines), change the slope, or influence both intercept and slope (unrelated lines).



Coincident Regression Lines

The simplest model is one in which the dummy variable has no effect on Y:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

In this case, the regression line is the same for both values of the dummy variable.

Parallel Regression Lines

Another model assumes that the dummy variable produces only an additive change in Y:

$$Y = \begin{cases} \beta_0 + \beta_1 x + \varepsilon, & d = 0 \\ \beta_0 + \beta_1 x + \beta_2 + \varepsilon, & d = 1 \end{cases}$$

Here, β_2 measures the additive change in Y due to the dummy variable.

Equal Intercepts, Different Slopes

A third model (the least common) assumes that the dummy variable changes the effect of x on Y:

$$Y = \begin{cases} \beta_0 + \beta_1 x + \varepsilon, & d = 0 \\ \beta_0 + (\beta_1 + \beta_3)x + \varepsilon, & d = 1 \end{cases}$$

Here, β_3 measures the change in the slope of the regression line due to d.

Unrelated Regression Lines

The most general model allows the dummy variable to affect both the intercept and slope:

$$Y = \begin{cases} \beta_0 + \beta_1 x + \varepsilon, & d = 0 \\ \beta_0 + \beta_1 x + \beta_2 + \beta_3 x + \varepsilon, & d = 1 \end{cases}$$

Equivalently,

$$Y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 (d \cdot x) + \varepsilon$$

In this case, β_2 measures the additive change in Y due to d, and β_3 measures the change in the effect of x on Y due to d.

How to Report and Test ANCOVA

Suppose the model is

$$Y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 (x \cdot d) + \varepsilon,$$

where x is a continuous covariate, $d \in \{0, 1\}$ is a dummy variable, and $x \cdot d$ is the interaction.

Step 1: Test the Interaction

- Fit the full model: $Y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 (x \cdot d) + \varepsilon$.
- Test $H_0 : \beta_3 = 0 \;\; v.s. \;\; H_a : \beta_3 \neq 0$.
  - If there is a significant interaction between x and d, interpret group-specific lines. And therefore the lines are unrelated (different intercepts and slopes).
  - If not significant, drop the interaction term and proceed to Step 2.

Step 2: Test the Main Effect of the Dummy Variable

- Fit the reduced model: $Y = \beta_0 + \beta_1 x + \beta_2 d + \varepsilon$.
- Test $H_0 : \beta_2 = 0 \;\; v.s. \;\; H_a : \beta_2 \neq 0$.
  - If significant, the regression line at each factor level has the same slope (β_1) but different intercept (β_0 v.s. $\beta_0 + \beta_2$). And therefore the lines are all parallel.
  - If not significant, drop the dummy variable and proceed to Step 3.

Step 3: Test the Main Effect of the Covariate x

- Fit the reduced model: $Y = \beta_0 + \beta_1 x + \varepsilon$.
- Test $H_0 : \beta_1 = 0 \;\; v.s. \;\; H_a : \beta_1 \neq 0$ to assess whether x is associated with Y. This is the same test from the previous section.

Example 1

`Salary1.csv` includes records of salary (in thousands of dollars), gender, and years of experience for each employee.

a. Fit a full model to predict the salary based on the gender and years of experience. Report the regression line.

b. Is there is a gender difference in salary after controlling for years of experience? Write down the pair of hypothesis and report the p-value.

c. Find the best model to predict the salary. Construct a plot of the fitted regression lines for males and females.

d. Using the best model, find the predicted salary for

   (i) A female with 5 years of experience

   (ii) A male with 5 years of experience

   (iii) A male with 10 years of experience

Example 2

Repeat Example 1 for `Salary2.csv` data.

Example 3

A small travel agency has retained your services to help them better understand two important customer segments. The first segment, denoted by A, consists of those customers who have purchased an adventure tour in the last twelve months. The second segment, denoted by C, consists of those customers who have purchased a cultural tour in the last twelve months. Data (`travel.csv`) are available on 925 customers. Note that the two segments are completely separate; there are no customers who belong to both segments. In addition, information is also available on the age of each customer, since age is thought to influence the amount spent.

a. Write down an appropriate regression model for the amount spent using both segment and age as predictors.

b. Find the best model to predict the amount spent. Construct a plot of the fitted regression lines for different segments.

c. Using the best model from the previous part, find the predicted amount for

   (i) A 40-year-old customer in segment A.

   (ii) A 40-year-old customer in segment C.

   (iii) A 50-year-old customer in segment C.

d. Based on your results, briefly comment on how segment and age influence customer spending.