Q(3) In a multilinear regression model;

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

where,

$$x_2 = \begin{cases} 1 & \text{if the province is over } 50\% \text{ Catholic} \\ 0 & \text{otherwise} \end{cases}$$

when both indicator variables x_2 and x_3 are included.

the matrix X:

4 observations

$$X = \begin{pmatrix} 1 & x_{11} & 1 & 0 \\ 1 & x_{22} & 1 & 0 \\ 1 & x_{33} & 0 & 1 \\ 1 & x_{44} & 0 & 1 \end{pmatrix}$$

Issue! Linear Dependence and Non-Invertibility

→ The issue here is that the columns in X, specially the intercept and the two indicator variables x_2 and x_3 are linearly dependent.

[which means that the sum of last two columns will become the intercept]

∴ The matrix X^TX becomes non-invertible

As we depend on invertion of X^TX to calculate least square estimates of the regression coefficients

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

→ The non-invertibility of X^TX prevents us from calculating the coefficients in a unique manner.

This linear dependence leads to perfect multicollinearity

where the predictors are not independent of each other

Multicolinearity consequences:

→ As of the perfect multicollinearity, the regression model cannot uniquely determine the coefficients β_2 (catholic) and β_3 (protestant).

→ The model can't differentiate between the effect of being Catholic or Protestant, as one predictor variable is simply a transformation of the other

→ which will lead to infinite number of possible solutions for the coefficient.

Solution: using c-1 indicator variables:

→ To resolve the above issue, we apply the principle that for any qualitative variable with c categories, we should use c-1 indicator variables.

→ In this model, as we have two cateogeries, we only need one variable

The regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

where,
$$x_2 = \begin{cases} 1 & \text{if the province is majority Catholic} \\ 0 & \text{otherwise} \end{cases}$$

→ The refference categories wold implicity represent majority Protestant provinces, and the effect of Protestantism would be captured by the intercept.

→ By dropping one of the indicator variables, we eliminate the issue of multicollinearity, ensuring that matrix $X^T X$ is now invertible.

→ This will allows us to obtain unique, interpretable estimates for the regression coefficients and accuratly measure the impact of being Catholic or Protestant relative to the baseline

\# This inclusion in both the indicator variables x_2 & x_3 cause perfect multicollinearity, making it impossible to uniquely estimate the coefficients due to the non-invertibility of $X^T X$. To address this, we must remove one of the indicator variables and represent the qualitative predictor with $c-1$ variables.

To avoid multicollinearity, we should use $c-1$ indicator variables, ensuring that the regression model is uniquely solvable

Q(5)

We are given a regression model that looks at the impact of temperature (x_1) and humidity (x_2) on growth rate (y) with an interaction term between these two variables:

The regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \cdot x_2) + \varepsilon$$

where:

y : plant growth rate (cm/day)

x_1 : Temperature (°C)

x_2 : Humidity (%)

$x_1 x_2$: Interaction term between temperature and humidity

ε : Error term

The estimated model is:

$$\hat{y} = 2 + 0.8x_1 + 0.5x_2 - 0.01(x_1 \cdot x_2)$$

(a) Without the interaction Term:

When we ignore the interaction Term, the model specifies to:

$$\hat{y} = 2 + 0.8x_1 + 0.5x_2$$

Here, we have to determine the change in plant growth when temperature increases by 1°C while holding humidity to a fixed level (60%)

→ The coefficient of x_1 is 0.8, which tells us that for each 1°C increase in temperature, plant growth increases 0.8 cm/day, assuming humidity remain constant.

→ In this simplified model, the relationship between temperature and plant growth in linear and independent of humidity.

∴ If humidity is held 60%. the change in plant growth rate for a 1°C increase in temperature is

$$\boxed{\Delta y = 0.8 \text{ cm/day}}$$

(b) With the interaction Term:

when we include the interaction term, the model becomes:

$$\hat{y} = 2 + 0.8x_1 + 0.5x_2 - 0.01(x_1 \cdot x_2)$$

Here, the interaction term $-0.01(x_1 \cdot x_2)$ suggests that the effect of temperature on plant growth depends on humidity. To find the effect of 1°C increase in temperature, we now need to calculate the partial derivative of $\hat{y}$ wrt x_1, accounting for the interaction with humidity (x_2):

$$\frac{\partial \hat{y}}{\partial x_1} = 0.8 - 0.01x_2$$

when humidity is fixed at 60%, we substitute $x_2 = 60$ into this equation

$$\frac{\partial \hat{y}}{\partial x_1} = 0.8 - 0.01 \times (60) = 0.8 - 0.6 = 0.2$$

Therefore, with the interaction term included, the average change in plant growth for a 1°C increase in temperature when humidity is 60% is :

$$\boxed{\Delta y = 0.2 \text{ cm/day}}$$

This interaction model highlights a complex relationship where the influence of temperature varies based on the humidity level. As humidity rises, the effect of temperature becomes less pronounced, demonstrating the importance of considering both factors together for accurate predictions.