Prof. Overbey Session 8 1/28/2025

Quantitative Reasoning II

Regression

What Is Regression?

Regression is a statistical (and machine learning) technique used to model and analyze the relationship between one variable (often called the *dependent variable* or *response variable*) and one or more other variables (often called *independent variables* or *explanatory variables*). The goal is typically to understand how changes in the explanatory variables are associated with changes in the response variable, and/or to make predictions for the response variable given values of the explanatory variables.

What Is Regression?

Question:
Which of the following scenario(s) are examples of regression?

A. Predicting the price of a house based on its size, location, and number of bedrooms.
B. Testing whether there is a difference in average exam scores between two groups of students using a t-test.
C. Modeling the relationship between a car's fuel efficiency (miles per gallon) and its weight, engine size, and age.
D. Determining whether a new drug significantly reduces blood pressure compared to a placebo using a randomized controlled trial.

What Is Regression?

Question:
Which of the following scenario(s) are examples of regression?

A. Predicting the price of a house based on its size, location, and number of bedrooms.
B. Testing whether there is a difference in average exam scores between two groups of students using a t-test.
C. Modeling the relationship between a car's fuel efficiency (miles per gallon) and its weight, engine size, and age.
D. Determining whether a new drug significantly reduces blood pressure compared to a placebo using a randomized controlled trial.

Adjustment

Which of the following best describes "adjustment" in the context of regression?

A. Including additional explanatory variables to account for potential confounding factors.

B. Changing the values of the response variable to fit the regression line.

C. Ensuring all explanatory variables have the same units before running the regression model.

D. Removing variables from the regression model that do not show a significant relationship with the response variable.

Adjustment

Which of the following best describes "adjustment" in the context of regression?

A. Including additional explanatory variables to account for potential confounding factors.

Correct. Adjustment means controlling for confounding variables by including them in the regression model, allowing us to isolate the effect of each variable on the response variable.

B. Changing the values of the response variable to fit the regression line.

Incorrect. The response variable is not modified; regression aims to model the relationship between the response and explanatory variables, not alter the data.

C. Ensuring all explanatory variables have the same units before running the regression model.

Incorrect. While ensuring consistent units can help interpretation, this is not the definition of adjustment.

D. Removing variables from the regression model that do not show a significant relationship with the response variable.

Incorrect. While removing insignificant variables can improve model simplicity, this is model selection, not adjustment.

Adjustment Explained

Adjustment in regression refers to the process by which we control or account for the effect of certain explanatory variables when examining the relationship of interest.

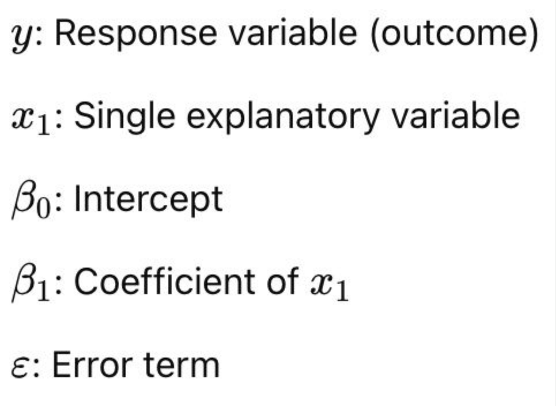
In other words, we “adjust for” variables that might confound (distort) the relationship between an explanatory variable of interest and the response variable.

- In **multiple regression**, you can include multiple explanatory variables to adjust for differences among observations that might otherwise bias the association you care about.
- By adjusting for **confounding variables**, you can isolate the “pure” effect of each explanatory variable on the response.

Linear Regression in Equations

In a **simple linear regression**, the equation is:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$



y: Response variable (outcome)

x_1: Single explanatory variable

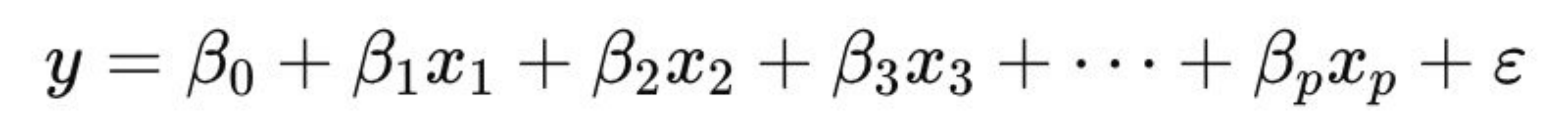
β_0: Intercept

β_1: Coefficient of x_1

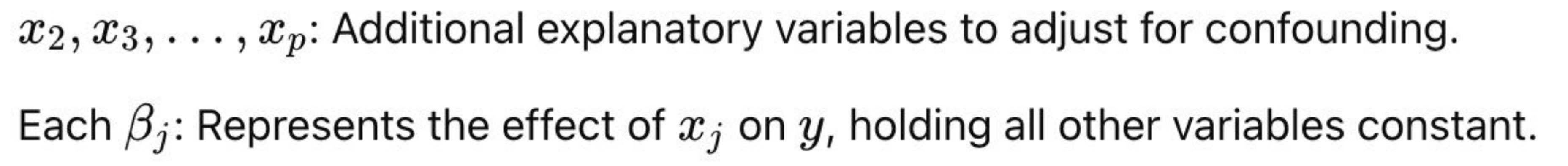
ε: Error term

Multiple Regression

Note: We have performed adjustment to include additional explanatory variables to control for potential confounders.



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p + \varepsilon$$



$x_2, x_3, \ldots, x_p$: Additional explanatory variables to adjust for confounding.

Each β_j: Represents the effect of x_j on y, holding all other variables constant.

Open Question

What are examples of models you can construct with multiple regression?

Open Question

What are examples of models you can construct with multiple regression?

Example: Predicting blood pressure (y) based on age (x1) and weight (x2)

$$\text{Blood Pressure} = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Weight}) + \varepsilon$$

Without adjustment (x2 omitted): The effect of age on blood pressure (β1) might be confounded by weight (e.g., older individuals tend to weigh more).

What Is the Error Term (ε)?

The **error term**, denoted by ε, represents the portion of the response variable (y) that the regression model cannot explain. It accounts for:

1. **Unobserved Factors**: Influences on y not captured by the explanatory variables in the model.
2. **Random Noise**: Random fluctuations or measurement errors in the data.
3. **Model Limitations**: Assumptions or simplifications in the model (e.g., assuming a linear relationship when the true relationship is more complex).

Error Term vs. Residual

While the terms "error term" and "residual" are closely related, they have subtle differences:

Error Term (ε):
Refers to the theoretical, unobservable differences between the actual y values and the true regression line in the population.

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

Residual (e):
Refers to the estimated differences between the observed y values and the predicted y values (from the fitted regression line in a sample).

$$e = y_{\text{observed}} - y_{\text{predicted}}$$