Class 14 Linear Regression for Causal Inference

Dr. Wei Miao

UCL School of Management

November 12, 2025

Learning Objectives

- Understand the basics of linear regression and its application in causal inference.
- Learn how to interpret regression coefficients and their significance.
- Gain hands-on experience in running and reporting regression analyses using R.

Section 1

Basics of Linear Regression

Linear Regression Models

- A simple linear regression is a model as follows.

$$y_i = \beta_0 + x_1\beta_1 + x_2\beta_2 + ... + x_k\beta_k + \epsilon_i$$

- y_i: Dependent variable/outcome variable
- x_k: Independent variable/explanatory variable/control variable
- β: Regression coefficients; β_0: intercept (should always be included)
- ϵ_i: Error term, which captures the deviation of Y from the line. Expected mean should be 0, i.e., $E[\epsilon|X] = 0$

Origin of the Name "Regression"

- The term "regression" was first coined by Francis Galton to describe a biological phenomenon: The heights of descendants of tall ancestors tend to regress down towards a normal average.
- The term "regression" was later extended by statisticians Udny Yule and Karl Pearson to a more general statistical context (Pearson, 1903).
- In supervised learning models, "regression" has a different meaning: when the outcome variable to be predicted is continuous, the task is called a regression task. This is because ML models are developed by computer science; causal inference models are developed by statisticians and economists.

Section 2

Estimation of Coefficients

How to Run Regression in R

- In this module, we will be using the `fixest` package, because it's able to accommodate more complex regressions, especially high-dimensional fixed effects.[1]

```
pacman::p_load(modelsummary, fixest)

OLS_result <- feols(
    fml = total_spending ~ Income, # Y ~ X
    data = data_full # dataset from M&S
)
```

[1] Fixed effects are a type of control variable that is constant within a group, such as country, year, or individual, to control for unobserved heterogeneity. See this link.

Report Regression Results

```
modelsummary(
    OLS_result,
    stars = TRUE # export statistical significance
)
```

	(1)
(Intercept)	-556.823***
	(21.654)
Income	0.022***
	(0.000)
Num.Obs.	2000
R2	0.629
R2 Adj.	0.629
AIC	29306.1
BIC	29317.3
RMSE	367.45
Std.Errors	IID

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Parameter Estimation: Univariate Regression Case

- Regressions with a single regressor are called univariate regressions. Let's take a **univariate regression** as an example:

$$total_spending = a + b \cdot income + \epsilon$$

- For each guess of a and b, we can compute the error for customer i,

$$e_i = total_spending_i - a - b \cdot income_i$$

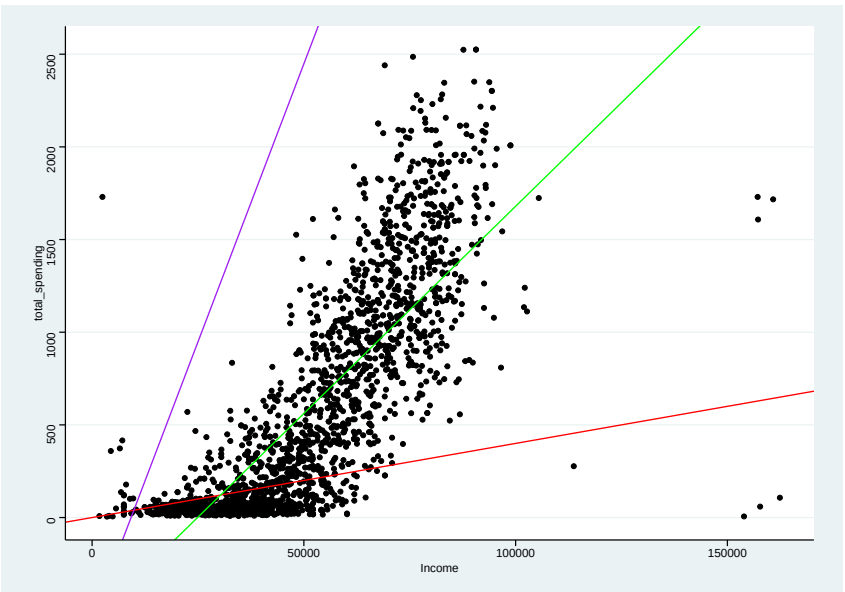
- We can compute the **sum of squared residuals (SSR)** across all customers

$$SSR = \sum_{i=1}^{n} (total_spending_i - a - b \cdot income_i)^2$$

- **Objective of estimation**: Search for the unique set of a and b that can minimise the SSR.
- This estimation method that minimizes SSR is called **Ordinary Least Square (OLS).**

Visualisation: Estimation of Univariate Regression

- If in the M&S dataset, if we regress **total spending** (Y) on **income** (X)



Multivariate Regression

- The OLS estimation also applies to multivariate regression with multiple regressors.

$$y_i = b_0 + b_1 x_1 + ... + b_k x_k + \epsilon_i$$

- **Objective of estimation**: Search for the **unique** set of b that can minimise the **sum of squared residuals**.

$$SSR = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_1 - ... - b_k x_k)^2$$

Section 3

Interpretation of Coefficients

Coefficient Interpretation

- Now on your Quarto document, let's run a new regression, where the DV is $total_spending$, and X includes $Income$ and $Kidhome$.

	(1)
(Intercept)	-299.119***
	(28.069)
Income	0.019***
	(0.000)
Kidhome	-230.610***
	(16.945)
Num.Obs.	2000
R2	0.661
R2 Adj.	0.660
AIC	29130.7
BIC	29147.5
RMSE	351.51
Std.Errors	IID

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

- **Controlling for** Kidhome, one unit increase in `Income` increases `total_spending` by £0.019.

Standard Errors and P-Values

- If we collect all data from the whole population, the regression coefficient is called the **population regression coefficient**.
- Because the regression is estimated on a random sample of the population, if we rerun the regression on different samples from the same population, we would obtain a different set of **sample regression coefficients** each time.
- In theory, the sample regression coefficient estimates follow a **t-distribution**: the mean is the true β. The **standard error** of the estimates is the estimated standard deviation of the error.
- Knowing that the coefficients follow a t-distribution, we can test whether the coefficients are statistically different from 0 using **hypothesis testing**.
- `Income`/`Kidhome` is statistically significant at the 1% level.

R-Squared

- R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by all included variables in a regression.
- Interpretation: 66% of the variation in `total_spending` can be explained by `Income` and `Kidhome`.
- As the number of variables increases, the R^2 will naturally increase, so sometimes we may need to penalise the number of variables using the so-called **adjusted R-squared**.

> **!** Important
>
> R-Squared is only important for supervised learning prediction tasks, because it measures the predictive power of the X. However, in causal inference tasks, R^2 does not matter much.