

Data Analysis in R

Multivariate Regression

Ken Stiller

17th January 2026

Syllabus: Data Analysis in R

1. Introduction
2. Causality & Basics of Statistics
3. Sampling & Visualisation
4. Prediction
5. **Multivariate Regression**
6. Probability & Uncertainty
7. Hypothesis Testing
8. Assumptions & Limits of OLS
9. Interactions & Non-Linear Effects

Table of Contents

Overview

OVB

Categorical Variables

Continuous Variables

Goodness of Fit

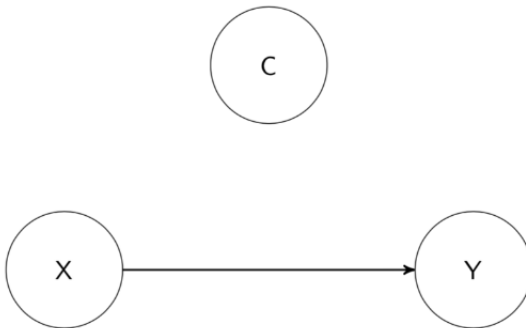
Wrap Up

Recap: Linear Regression Analysis

- ▶ We want to make predictions minimising errors
- ▶ We could simply use the mean of a variable
- ▶ If we have another variable that we suspect may be associated with the variable we care about, we can use linear regression to help make better predictions
- ▶ A linear regression model is a *linear* approximation of the relationship between explanatory variables (e.g., X) and a dependent variable Y
- ▶ We do this by minimising the sum of squared errors

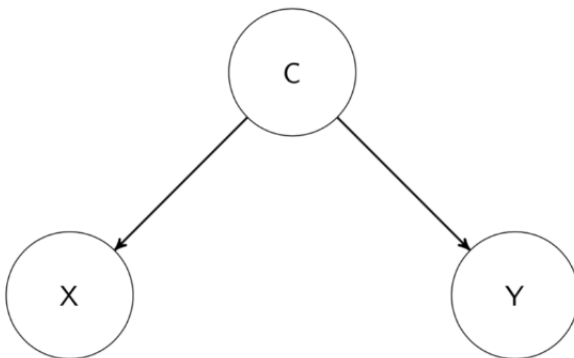
Omitted Variable Bias

Omitted Control



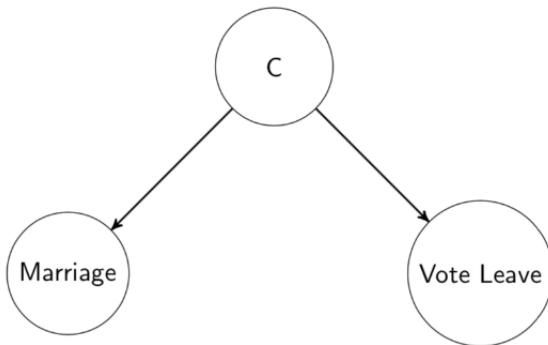
Omitted Variable Bias II

Omitted Control



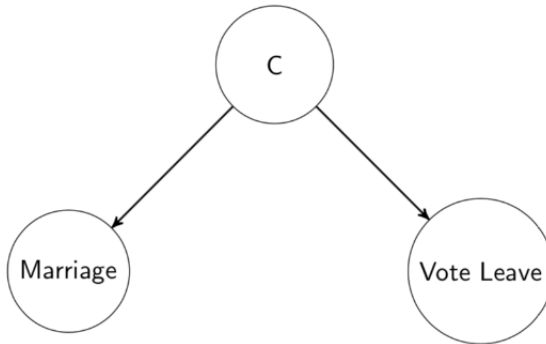
Omitted Variable Bias III

What is a potential omitted control?

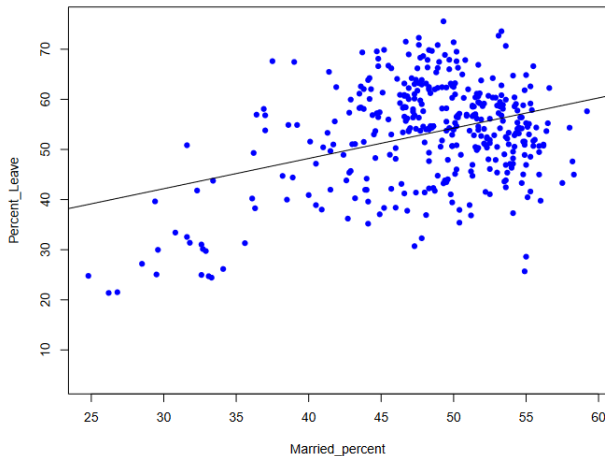


Omitted Variable Bias IV

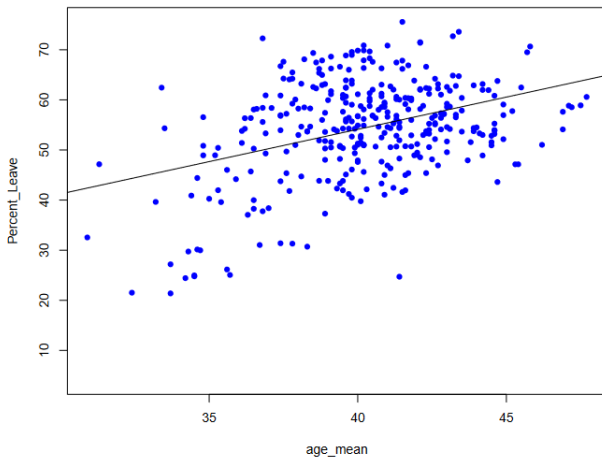
Omitted control: age



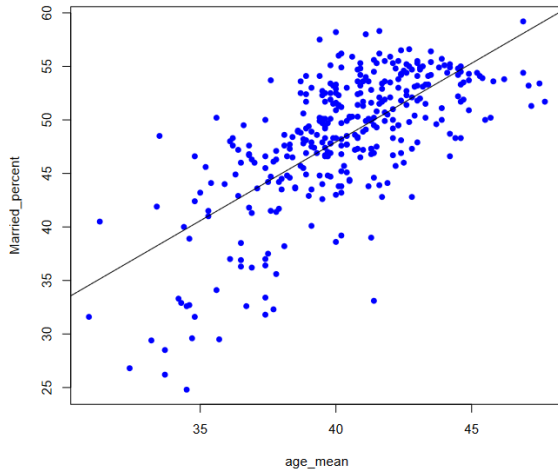
Mutual Association: Scatter-plot of % Leave and % Married



Mutual Association: Scatter-plot of % Leave and Mean Age



Mutual Association: Scatter-plot of % Married and Mean Age



Multivariate OLS: Interpretation of Coefficients

Table: Effect of % of Marriage on %Vote Leave

	(1) % Vote Leave	(2) % Vote Leave
% Married ($\hat{\beta}_1$)	0.604	0.402
Average age ($\hat{\beta}_2$)		0.697
Constant ($\hat{\alpha}$)	24.062	6.988
Observations	2,152	2,152
R Squared	0.1296	0.181

Source: British election survey 2017

- (1) A one unit increase of % Married is associated with 0.604 increase in % Vote Leave
- (2) A one unit increase of % Married is associated with 0.402 increase in % Vote Leave **holding Average Age constant**
- In (2) $\hat{\beta}_1$ estimates the **partial effect** of X_1 on Y

The Logic of Multivariate Regression

- ▶ In order for $\hat{\beta}$ to be unbiased we need the following condition:
 $E[\epsilon|X_1] = 0$ (**zero conditional mean**)
 - ▶ In observational studies, X_1 is likely to be determined by omitted variables in ϵ , which could be also related to Y
 - ▶ thus, $E[\hat{\beta}] \neq \beta$
 - ▶ This is known as **omitted variable bias**
- ▶ A common practice that aims to account for omitted variable bias is to use X_2 (the confounder) as a '*control*':

$$Y = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1}_{\text{slope 1}} X_1 + \underbrace{\beta_2}_{\text{slope2}} X_2 + \underbrace{\epsilon}_{\text{error term}}$$

previous error term

- ▶ Holding X_2 constant, β_1 denotes the partial **association** of X_1 with Y
- ▶ β_0 now denotes the expected value when **all independent variables** are 0 (whether this is useful or not)

General Extension of Multivariate Regression

- ▶ A multiple variable regression can be written as:

$$Y = \alpha + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon_i$$

- ▶ X_k are the k independent variables and β_k are the k coefficients for those X_k variables
- ▶ How many and which ones can [**should**] be added?
 - ▶ If the intention is to estimate the effect of X_1 on Y , controls *can* be useful in addressing omitted variable bias
 - ▶ Associated **both** with X_1 and Y
 - ▶ Number is always limited by the degrees of freedom $(N - k)$, where N is the number of observations
 - ▶ Our aim is to meet the **conditional independence assumption** ($E[\epsilon | X_1, \dots, X_k] = 0$)

A brief example

Regression Results with Dummy Variables

- ▶ Consider the example of colonial legacy on democratisation
- ▶ $[Democracy|Colony] = \beta_0 + \beta_1 Colony$
- ▶ What is β_0 here?
 - ▶ β_0 : The mean level of Democracy for **non-colonies**
- ▶ What about β_1
 - ▶ β_1 : The difference in the level of Democracy **between colonies and non-colonies**.
- ▶ *Question*: How do we know the level of Democracy for **colonies**?
- ▶ Now imagine, we want to estimate
 $E[Democracy|GDP] = \beta_0 + \beta_1 GDP$.

How does **colony** play into this?

Adding Categorical Covariates

- We can generalize the prediction equation:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

- This implies that we want to predict Y using the information we have about X_1 and X_2
- Therefore:

$$Democracy = \hat{\beta}_0 + \hat{\beta}_1 GDP + \hat{\beta}_2 Colony$$

What Does It Mean to Add Covariates?

- ▶ Colony is a *dummy variable*. It takes only two values:
 - ▶ 0 if the country **was not** a British colony
 - ▶ 1 if the country **was** a British colony
- ▶ Based on our regression equation, this renders two regression lines over GDP:
 - ▶ If $X_2 = 0$: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 * 0 = \hat{\beta}_0 + \hat{\beta}_1 X_{1i}$
 - ▶ If $X_2 = 1$:
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 * 1 = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 X_{1i}$$
- ▶ We are fitting two lines with the **same slope** but **two different intercepts**
 - ▶ Think of it as adding a constant to former British colonies

Where's the Difference?

From R, we get the following estimates: $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$:

FHREVERS	Coef.
GDP90LGN	1.705888
BRITCOL	.5880665
_cons	-1.506045

Non-British Colonies:

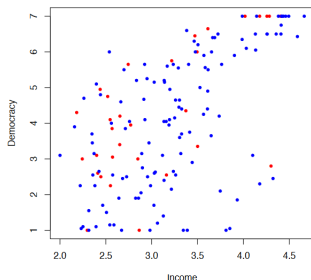
$$\blacktriangleright \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

$$\blacktriangleright \hat{Y} = -1.5 + 1.7 * X_{1i}$$

Former British Colonies:

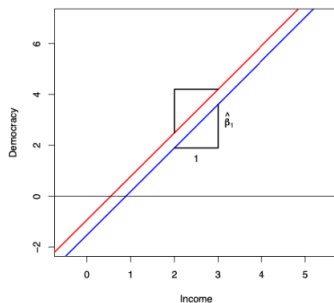
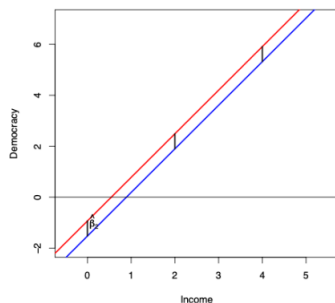
$$\blacktriangleright \hat{Y} = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 X_{1i}$$

$$\blacktriangleright \hat{Y} = -.92 + 1.7 * X_{1i}$$



Different Intercept - Same Slopes

Same slope, but two different intercepts. Two different levels, corresponding to the different values in X_2 : colony.



Recall: Reference Categories

- ▶ Imagine we have the following set-up:
- ▶ Our **dependent variable** (Y) is **life satisfaction** (0 to 10 ordinal scale)
- ▶ Our **independent variable** (X) is **civil status**
 - ▶ $X = 1$ if married
 - ▶ $X = 2$ if divorced
 - ▶ $X = 3$ if widowed
 - ▶ $X = 4$ if single
- ▶ Can a regression coefficient be interpreted with the variable coded like this?

Recall: Reference Categories II

- For us to make sense of the results, we recode the categorical variable into a set of dummies:

$$LifeSatisfaction = \beta_0 + \beta_1 Divorced + \beta_2 Widowed + \beta_3 Single + \mu_i$$

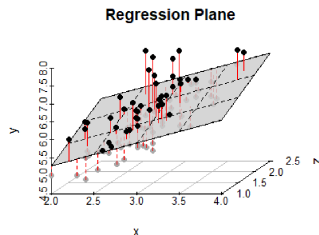
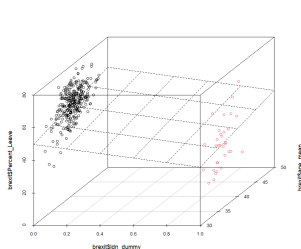
	Estimate
divorced	-0.605
widowed	-0.939
single	0.220
constant	6.640

- Categorical variables (dummies or variables with more than 2 categories) are always come with a reference category
- The coefficients are all interpreted **with reference to this category**

Ultimately, which category you pick as reference category does not matter in statistical terms

Multivariate Regression with Continuous Variables

- ▶ We also include continuous variables as controls - the basic logic is exactly the same
- ▶ Effects are also interpreted in the same way - but this is less intuitive than with categorical variables
- ▶ Regression with two continuous variables fits a **plane** (*not a line*) made up of two perpendicular dimensions
- ▶ We are still trying to reduce squared errors



Worth It? Goodness of Fit Revised

- ▶ The *R-squared* remains a commonly used way of assessing goodness of fit
- ▶ The way you calculate the *R-squared* in the multivariate context is exactly the same as in the bivariate context
- ▶ If we keep adding variables, the *R-squared* will increase by design
- ▶ To keep it from doing so mechanically, we usually rely on the **adjusted R-squared**:

$$R^2_{adjusted} = \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where p = number of predictors

Some Words of Caution...

- ▶ Controls *can* be useful - but they are not a magical solution
- ▶ Their assumptions are quite (prohibitively?) strong: We'd have to think of and measure *all* confounders to estimate the unbiased effect
 - ▶ One can usually think of some additional confounders
 - ▶ Often, confounders are hard to measure (e.g., charisma in election campaigns) or unobservable entirely
- ▶ Be careful about what controls you choose - controlling for anything that can be a **consequence of X** reintroduces bias!
 - ▶ This is called **post-treatment bias**
 - ▶ All controls must be pre-treatment, i.e. realised before X
 - ▶ We also don't want to include more than one variable measure conceptually the same thing or even be perfectly collinear

Key Take Aways

- ▶ In social sciences, anything is related to virtually everything else, so there are many confounders
- ▶ Controlling them 'away' is a common approach to tackle the issues, but it comes with *[too]* strong assumptions
- ▶ Statistical interpretations of multivariate regressions are based on the *ceteris paribus* assumption
- ▶ Be aware of substantive issues when analysing dummy/categorical variables
- ▶ We got an idea of how continuous controls work
- ▶ Don't use controls that are realized after your independent variable (post-treatment)
- ▶ Remember to be skeptical of R-squared and the substantive meaning of adjusted R-squared

The Way Ahead...

- ▶ So far we have been concerned with point estimates: what is the effect of X on Y ?
- ▶ Next week, we'll move from how we find a single answer to understanding how **precise/uncertain** this answer is
- ▶ Bear in mind throughout: we still care about how "large" an effect is. But we also want to know if our estimate of that effect is precise or not. **Both** are crucial to interpret a result and provide substantive answers to research questions