Applied Statistical Analysis I

Multiple linear regression

Elena Karagianni, PhD Candidate
karagiae@tcd.ie

November 5, 2025

Department of Political Science, Trinity College Dublin

Multiple linear regression

Why do we need multiple linear regression?

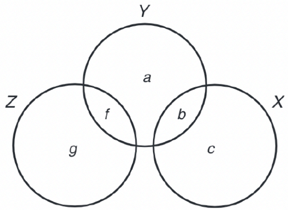


Figure 9.2. Venn diagram in which X and Z are correlated with Y, but not with each other.

"In that case — which, we have noted, is unlikely in applied research — we can safely omit consideration of Z when considering the effects of X on Y. In that figure, the relationship between X and Y (the area b) is unaffected by the presence (or absence) of Z in the model."

Kellstedt and Whitten (2018, 213)

Why do we need multiple linear regression?

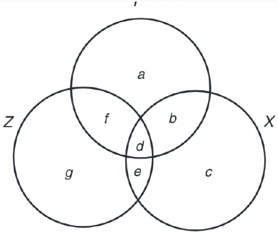


Figure 9.1. Venn diagram in which X, Y, and Z are correlated.

"If, hypothetically, we erased the circle for Z from the figure, we would (incorrectly) attribute all of the area $b + d$ to X, when in fact the d portion of the variation in Y is shared by both X and Z. This is why, when Z is related to both X and Y, if we fail to control for Z, we will end up with biased estimates of X's effect on Y"

Kellstedt and Whitten (2018, 212)

Adding Covariates

Multiple regression analysis allows us to add covariates $X_2, \ldots, X_k$ on top of X_1 in a regression of Y:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

Adding Covariates

Multiple regression analysis allows us to add covariates $X_2, \ldots, X_k$ on top of X_1 in a regression of Y:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

- A *bivariate* regression of Y on X_1 does not yield accurate predictions of Y. We need additional covariates to minimize the prediction error $(\hat{Y} - Y)$.

Adding Covariates

Multiple regression analysis allows us to add covariates $X_2, \ldots, X_k$ on top of X_1 in a regression of Y:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

- A *bivariate* regression of Y on X_1 does not yield accurate predictions of Y. We need additional covariates to minimize the prediction error $(\hat{Y} - Y)$.
  - If we were talking about *causality*, we would say that a bivariate regression of Y on X_1 does not yield an unbiased estimate of the true effect β_1 of X_1 on Y. We need to adjust for additional covariates to minimize the bias $(\hat{\beta}_1 - \beta_1)$.

Adding Covariates

Multiple regression analysis allows us to add covariates $X_2, \ldots, X_k$ on top of X_1 in a regression of Y:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

- A *bivariate* regression of Y on X_1 does not yield accurate predictions of Y. We need additional covariates to minimize the prediction error $(\hat{Y} - Y)$.
    - If we were talking about *causality*, we would say that a bivariate regression of Y on X_1 does not yield an unbiased estimate of the true effect β_1 of X_1 on Y. We need to adjust for additional covariates to minimize the bias $(\hat{\beta}_1 - \beta_1)$.
- The estimation is the same for both purposes; it is the rationale underlying model specification that changes.

Multiple Linear Regression Model

- The general multiple linear regression model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

  - The error term, ϵ, contains factors other than $X_1, \ldots, X_k$ that affect Y. We assume that all factors in the unobserved error term are uncorrelated with the explanatiry variables.
- The estimation approach is the same as in the two-variable case, i.e., to minimize the sum of squared residuals (but now we get $k+1$ normal equations):

$$\min_{\hat{\alpha},\ldots,\hat{\beta}_k} \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_k X_{ik})$$

Interpreting the Coefficients

- Interpretation of regression coefficients:

  $\hat{\alpha}$ = Predicted value of Y when all X's equal zero.

  $\hat{\beta}_1$ = On average, a one-unit change in X_1 leads to a $\hat{\beta}_1$-unit change in Y, holding everything else constant.

  $\vdots$

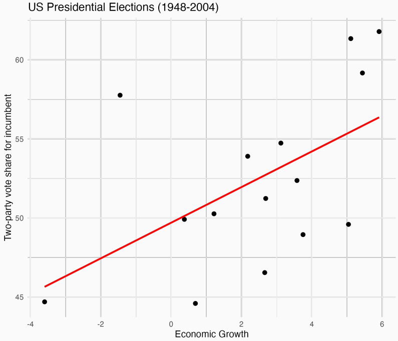
  $\hat{\beta}_k$ = On average, a one-unit change in X_k leads to a $\hat{\beta}_k$-unit change in Y, holding everything else constant.

- Note that in $k+1$-dimensional space, a fitted multiple regression model no longer defines a line, but a *hyperplane*.
- For $k=2$, OLS means fitting a least squares plane that best fits the cloud of data points in a three-dimensional space.

An example

An example

- Let's look at the US presidential election example again.
- We used economic growth to predict the two-party vote share:



- Can you think of an alternative explanation for success of the presidential party?

An example

- Let's consider presidential popularity in addition to economic growth.
- Popularity for presidents prior to election ranged between 31% and 74%.
- Our model becomes:

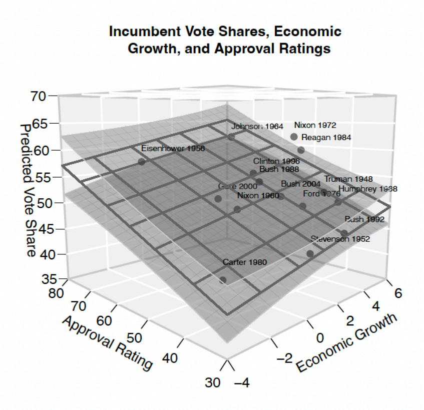
$$\text{VoteShare} = \alpha + \beta_1 \text{Growth} + \beta_2 \text{Approval} + \epsilon$$

- Results of the regression of vote share on growth and approval rating:

Variable	Estimate	SE
Constant	34.83	2.77
Growth	0.81	0.27
Approval	0.32	0.06
R^2 = 0.81		
Obs. = 15		

Visualizatin of example

$$\widehat{\text{VoteShare}} = 34.83 + 0.81 \times \text{Growth} + 0.32 \times \text{Approval}$$



Statistical Control

Statistical Control

So what is *statistical control* and how do we get an effect of of ‘Approval’ of 0.32 independent of ‘Growth’?

Statistical Control

- We control for ‘Growth’ by removing its effect from ‘VoteShare’ and ‘Approval’ before regressing them.

Statistical Control

- We control for 'Growth' by removing its effect from 'VoteShare' and 'Approval' before regressing them.
    - What is the effect of approval rating on vote share given growth?
        1. Let's start by running the simple regression Vote Share $\sim$ Growth

DV: Vote Share	Estimate	SE
Constant	49.70	1.75
Growth	1.13	0.49

    - The residuals in this model are the part of 'Vote Share' unexplained by 'Growth'
    - In other words, we remove the effect of growth from the vote share variable.

Statistical Control

- Now, we remove the effect of 'Growth' (X_1) from 'Approval' (X_2)
    2. Let's regress Approval $\sim$ Growth

DV: Approval	Estimate	SE
Constant	46.54	4.67
Growth	0.98	1.32

- The residuals in this model are the part of 'Approval' unexplained by 'Growth'
- In other words, we remove the effect of growth from the approval rating.

Statistical Control

3. Now, we regress the residuals from step (1) on the residuals from step (2):

DV: Residuals 1	Estimate	SE
Constant	0.00	0.66
Residuals 2	0.32	0.06

- Having controlled for the effect of Growth (removed from both vote share and approval), we can compute the **unconfounded (independent) effect of approval** on vote share.

Statistical Control

3. Now, we regress the residuals from step (1) on the residuals from step (2):

DV: Residuals 1	Estimate	SE
Constant	0.00	0.66
Residuals 2	0.32	0.06

- Having controlled for the effect of Growth (removed from both vote share and approval), we can compute the **unconfounded (independent) effect of approval** on vote share.
- This procedure gives exactly the same coefficient (and SE) as a multiple regression:

DV: VoteShare	Estimate	SE
Constant	34.83	2.77
Growth	0.81	0.27
Approval	0.32	0.06

Statistical Control

3. Now, we regress the residuals from step (1) on the residuals from step (2):

DV: Residuals 1	Estimate	SE
Constant	0.00	0.66
Residuals 2	0.32	0.06

- Having controlled for the effect of Growth (removed from both vote share and approval), we can compute the **unconfounded (independent) effect of approval** on vote share.
- This procedure gives exactly the same coefficient (and SE) as a multiple regression:

DV: VoteShare	Estimate	SE
Constant	34.83	2.77
Growth	0.81	0.27
Approval	0.32	0.06

- Multiple regression coefficients can statistically control for (be independent of) the effects of other variables.