A Practical Review of Multiple Linear Regression

Advanced Topics in Quant Social Research

Plan for the day

- Recap: linear regression model
  - Definition and setup
  - Data analytical objectives
  - Assumptions and inference
- "Multiple" linear regression
  - Why "multiple"
  - Principles of model specification
  - Practical reminders
- Looking ahead: "All models are wrong, but some are useful" (Box and Draper, 1987)

What is linear regression model

- Statistical inference aims to understand the **relationship** between different **variables**.

- The linear regression model is a common technique to use a **linear function** to represent and **estimate** the statistical relationship between **X** and **Y**, using the data we have.

$$\mathbf{Y} = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} \mathbf{X} + \underbrace{\epsilon}_{\text{error or disturbance}},$$

where

- **Y** is the outcome or response variable
- **X** is the predictor or independent variable

What is linear regression model

$$\mathbf{Y} = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} \mathbf{X} + \underbrace{\epsilon}_{\text{error}},$$

where

- **Intercept**: The corresponding value of $\mathbf{Y}$ when $\mathbf{X}$ is set at 0.
- **Slope**: The corresponding change in the value of $\mathbf{Y}$ when we increase $\mathbf{X}$ by one unit.
- **Error** or **disturbance**: The corresponding deviation in the value of $\mathbf{Y}$ from the predicted $\widehat{\mathbf{Y}}$ as follows:

$$\begin{aligned} \epsilon &= \text{actual outcome} - \text{predicted outcome} \\ &= \mathbf{Y} - \widehat{\mathbf{Y}} \\ &= \mathbf{Y} - (\alpha + \beta\mathbf{X}) \end{aligned} \tag{1}$$

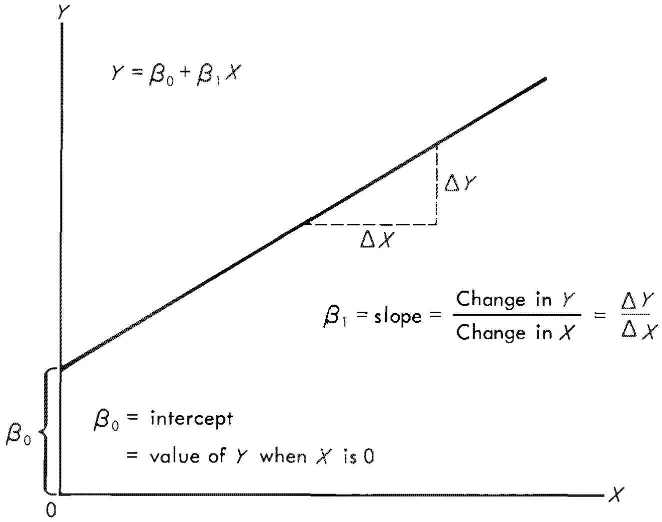
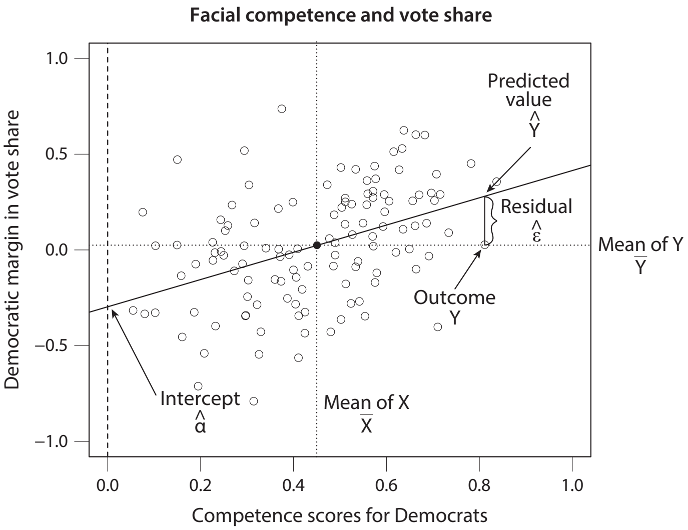


FIGURE 3-1 Equation of a straight line


Facial competence and vote share
Democratic margin in vote share
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0
Competence scores for Democrats
Predicted value $\hat{Y}$
Residual $\hat{\varepsilon}$
Mean of Y $\bar{Y}$
Outcome Y
Intercept $\hat{\alpha}$
Mean of X $\bar{X}$

Different views of linear regression model

- We use the linear regression model to estimate or quantify the statistical relationship, or **correlation**, between **X** and **Y**.

- Depending on different data analytical objectives, **X** can be understood in different ways (BdM and Fowler 2021).

  - Regression for **explanation** is to use **X** (as explanatory variable) to explain the variation in **Y**.

  - Regression for **forecasting** is to use **X** (as predictor) to predict **Y**.

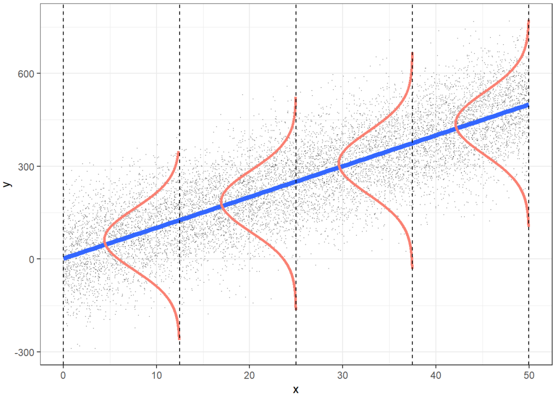
- Regression for **causal inference**: With certain assumptions (more after Week 5), the estimated slope (or $\widehat{\beta}$) can be considered as the **marginal effect** of **X** (as cause or treatment).

Statistical inference of linear regression

- For the linear regression model to produce **unbiased** $\widehat{\beta}$ (i.e., the estimated slope is the true slope), we need several assumptions or conditions (Roback and Legler 2020).
  - Linearity: There is a **linear** relationship between $\mathbf{X}$ and $\mathbf{Y}$.
  - Independence: The errors of individual observations (i.e., $\mathbf{Y} - \widehat{\mathbf{Y}}$) are **independent** of each other.
  - Normality: Across different values/levels of $\mathbf{X}$, $\mathbf{Y}$ is **normally distributed**.
  - Homoscedasticity: Across different values/levels of $\mathbf{X}$, the variance or standard error of $\mathbf{Y}$ is equal
- **[Note]** These assumptions do not say the linear regression model will produce **efficient** $\widehat{\beta}$ (i.e., the estimated slope may still have large variance or standard error).

Statistical inference of linear regression

We can combine the **independence**, **normality** and **homoscedasticity** assumptions and re-state them: **Across different values/levels of X, Y should be independent and identically (and normally) distributed**.



Statistical inference of linear regression

- Many statistical tools have been developed to detect and evaluate assumption violations.

- The violation of these assumptions is by no means a deal break, as many statistical tools or alternative models/regression estimators have been developed to address these situations.
  - **Generalized linear models** are developed to get round the linearity and normality assumptions (more in Weeks 3-4).
  - **Time-series analysis** is a well-known technique to tackle the violation of the independence assumption (not covered).
  - **Robust** or **clustered** standard errors are developed to address the violation of the homoscedasticity assumption (not covered).
  - It is also possible to ditch the linear regression and, instead, use **non-parametric** or **Bayesian** statistical analysis (more in Week 11).

Statistical inference of linear regression

- Is β "statistically" significant? When using the linear model to estimate β, we we need to evaluate two hypotheses:

$$\begin{aligned} H_0 &: \beta = 0 \\ H_1 &: \beta \neq 0 \end{aligned} \tag{2}$$

- The **null** hypotheses (H_0) means **X** and **Y** are not correlated while the **alternative** hypotheses (H_1) says the opposite.
- To show correlation, we need to reject H_0.
  - The p-value: The conditional probability of observing our data given H_0 (the probability should be low enough for us to reject H_0)
  - The confidence interval: The possible range of our $\widehat{\beta}$ (the range should not include 0 for us to reject H_0)
- In addition to **statistical** significance, we should also check the sign and size of $\widehat{\beta}$.

Statistical inference of linear regression

- Is the proposed model specification the best **fit**? That is, does the model provide $\widehat{\mathbf{Y}}$ with smallest residuals?
  - R^2: The idea is to **maximize** the proportion of variance of $\mathbf{Y}$ explained by the variance of $\widehat{\mathbf{Y}}$.

$$R^2 = \frac{\text{variance of } \widehat{\mathbf{Y}}}{\text{variance of } \mathbf{Y}} \tag{3}$$

  - Sum of squared residuals (SSR): The idea of **least squares** is to **minimize** the SSR.

$$\mathbf{SSR} = \sum_{i=1}^{n} \widehat{\epsilon}_i^2 = \underbrace{\widehat{\epsilon}_1^2 + \widehat{\epsilon}_2^2 + \widehat{\epsilon}_3^2 + \widehat{\epsilon}_4^2 + \widehat{\epsilon}_5^2 \cdots + \widehat{\epsilon}_{n-1}^2 + \widehat{\epsilon}_n^2}_{\text{sum of the squared estimated errors of } n \text{ observations}}, \tag{4}$$

where n refers to the number of observations in the model.

"Multiple" linear regression

- Given the complexity of the social world, it is impossible to use a single **X** to predict or explain the outcome of interest.

- The **Multiple** (or **multivariate**) linear model is the most widely used linear regression model where we include more than one **X**.

- If we use k to refer to the number of **X**s in the model, a multiple linear regression model can be represented as follows

$$\mathbf{Y} = \alpha + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \cdots + \beta_k \mathbf{X}_k + \epsilon, \tag{5}$$

  where k refers to the kth **X**.

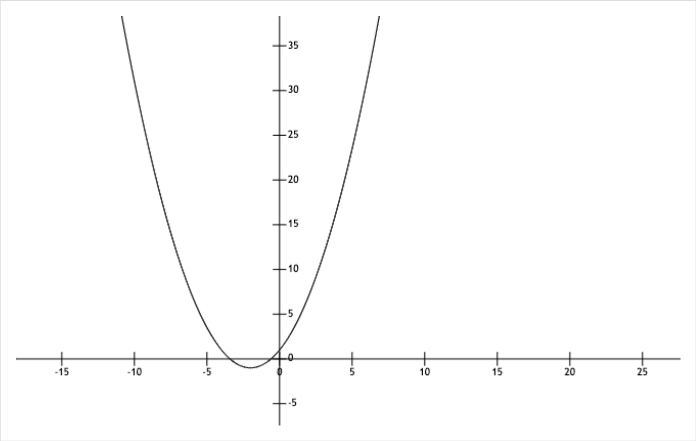
“Multiple” linear regression

$$\mathbf{Y} = \alpha + \widehat{\beta}_1 \mathbf{X}_1 + \cdots + \widehat{\beta}_k \mathbf{X}_k$$

- In the model above, $\widehat{\beta}_i$ refers to the slope of a particular X_i.
- The estimated slope of X_i, $\widehat{\beta}_i$, can describe the correlation between $\mathbf{Y}$ and $\mathbf{X}_i$ when we **control for** other predictors in the model.
- Likewise, there can be different interpretations:
  - β_i shows the correlation between $\mathbf{X}_i$ and $\mathbf{Y}$ while accounting for the association between other $\mathbf{X}$s and $\mathbf{Y}$.
  - β_i shows the causal effect of $\mathbf{X}_i$ on $\mathbf{Y}$ while ruling out the effect of other $\mathbf{X}$s on $\mathbf{Y}$ (with certain assumptions).

“Multiple” linear regression

- When we include multiple **X**s in our model,
  - We are less likely to commit the **omitted variable bias**, but
  - We need to check for **collinearity** (using **variance inflation factor**) and **overfitting** (using **adjusted** R^2).
- We can also consider more complicated statistical associations between **Y** and different **X** by including
  - Interaction term (e.g., $\mathbf{Y} = \alpha + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_1 \mathbf{X}_2 + \epsilon$)
  - Quadratic or polynomial terms (e.g., $\mathbf{Y} = \alpha + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_1^2 + \epsilon$)
- Model comparison is recommended but not essential.



$$\mathbf{Y} = 1 + 2\mathbf{X} + 0.5\mathbf{X}^2 \tag{6}$$

Conclusion: "All models are wrong, but some are useful"

- Practical reminders to build your multiple linear regression model,
  - Drawing on your knowledge of the subject matter, **specify a baseline model** including all key explanatory variables to avoid **omitted** variable bias.
  - Check if the baseline model falls into the victim of **collinearity** and/or **overfitting**.
  - Consider more complicated model specification techniques, such as interaction and/or quadratic/polynomial terms.
  - Consider alternative estimators or modeling choices, such as generalized linear model (Week 3) and/or multilevel/hierarchical modeling (Week 4)
- **All models are approximations**. Good analysis requires careful decision-making while holding a solid knowledge of the subject matter.

Key texts

- *Quantitative Social Science: An Introduction* (Imai 2018), Chapter 4
- *Beyond Multiple Linear Regression* (Roback and Legler 2020), Chapter 1
- *Thinking Clearly With Data* (BdM and Fowler 2022), Chapter 2
- *The Effect* (Huntington-Klein 2022), Chapters 4 and 13

Example: Fitting OLS to 4 Observations