Tutorial: Multiple Linear Regression

Hammad Shaikh

October 6, 2019

Fitness of Regression Model

R^2 relevant for prediction but not param inference

- R^2 measures the proportion of variation in the outcome (Y) explained by the independent variable(s) (X)
  - R^2 is a number between 0 and 1

    ↳ $R^2 \approx 1$ ⇒ model fits data well

- $R^2 = \frac{SSR}{SST}$; SST = Sum of Square Total, SSR = Sum of Square Regression
  - SST $= \sum_{i=1}^{n}(y_i - \bar{y})^2$ and SSR $= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

- R^2 applies to both simple and multiple linear regression

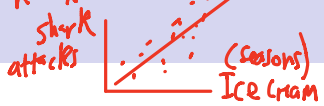
Simple Linear Regression Summary

- The population linear regression model
  - $Y = \beta_0 + \beta_1 X + \epsilon$
    ↳ paran: β_0, β_1
- Line of best fit and OLS estimator
  - $\widehat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)}$ and $\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$ $\left(\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X\right)$
- Hypothesis testing
  - $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$ $\Rightarrow$ pvalue $< \alpha$

X does not impact Y
- Measures of fit for simple regression: $\widehat{y} = \widehat{b}_0 + \widehat{b}_1 x$
  - Correlation and R^2

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_R X_R + \varepsilon$

Extending to Multiple Regression



- Results from simple linear regression are usually not causal
  - Other factors that affect both X and Y are not considered
    - Can bias slope estimates (omitted variable bias)

Y = Shark attacks, X = Ice cream sold $\Rightarrow \hat{\beta}_1 >> 0$

- Returns to education: $AdultIncome_i = \beta_0 + \beta_1 YrsEduc_i + \epsilon_i$
  - What are some variables in ϵ_i that may bias $\widehat{b}_1$?

$X \longrightarrow Y \Rightarrow Cov(X, \varepsilon) \neq 0$ (Bias)

ε: {effort, ability, motivation}

- Two solutions to help obtain causal result:
  - 1) Randomized control trial, or 2) Multiple regression

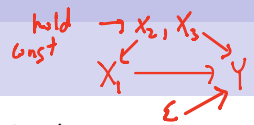
↳ $X \perp \varepsilon$

↳ Hold observed factors constant

Randomized Control Trial

- Simple regression model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- In a RCT the Xs are randomly assigned to individuals
  - No omitted variable bias since X_i independent to ϵ_i
    - Now $\widehat{b}_1$ has a causal interpretation
- Correlation does not imply causation?
  - Generally true for observational data, but false for experimental data where treatment variable is randomly assigned
- Returns of education: $Y_i = \beta_0 + \beta_1 YrsEduc_i + \epsilon_i$
  - Can we randomly assign years of education to individuals?

Multiple Regression



- Slope estimate in simple regression can be biased from omitted variables related to X and Y
  - Solution is to include the omitted variables into the model

- Multiple regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$
  - β_1 = effect of changing X_1 on Y holding $X_2, \ldots, X_k$ constant
  - $\widehat{b}_1$ can be causal if all relevant variables are included
    - Conditional independence: ϵ indep. to X_1 given $X_2, \ldots, X_k$

↳ $\varepsilon \perp X_1 \mid X_2, \ldots, X_R$ constant

- Returns to education:
  $Y_i = \beta_0 + \beta_1 YrsEduc_i + \beta_2 Exp_i + \beta_3 ParentIncome_i + \epsilon_i$ (motivation)

↳ $\hat{\beta}_1$ is change in Y wtn ↑Yrs Educ holding constant Exp, Parent Inc.

Regression Table Example

① & ③ are simple linear reg
② & ④ are multiple reg.

Table: Income and Health Returns to Education (Fake Data)

	(1) Hourly Wage	(2) Hourly Wage	(3) Years Lived	(4) Years Lived
Constant	11***	10***	65***	66***
	(2.5)	(0.1)	(10)	(10)
Years of Educ	2***	1***	2***	3***
	(0.5)	(0.1)	(0.25)	(0.3)
Experience		3***		0.5**
		(0.8)		(0.245)
Parent income ($1000)		0.1**		0.15*
		(0.048)		(0.075)
R-square	0.15	0.30	0.10	0.20
No. of indivisuals	15000	15000	15000	15000

Stars denote level of significance *10%,** 5%, and ***1%.

X

Y

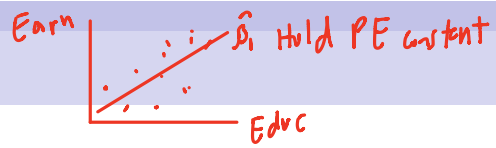
SE

t = 4

N = 15000

- Regression table generally contain coefficient estimates, standard errors, no. of observations, and R^2

"Partialing Out" Idea



$\hat{\beta}_1$ Hold PE constant

- Suppose we want to make scatter plot of earnings on education but adjust for parental education

$$Earn = \beta_0 + \beta_1 Educ + \beta_2 PE + \varepsilon$$

- Step 1: We need to obtain variation in education that is independant of the parental education

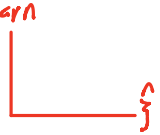
$$\hookrightarrow Educ = \alpha_0 + \alpha_1 PE + \hat{\xi} = \hat{\xi} \text{ (residuals)}$$

Var. in Educ. not explained by PE

- Step 2: Related earnings on the variation obtained in step 1.

$$\hookrightarrow Earn = \gamma_0 + \gamma_1 \hat{\xi} + \mu$$

$$\hookrightarrow \hat{\gamma}_1 = \hat{\beta}_1$$



Summary of Linear Regression

- Goal: examine causal relationship between outcome Y and explanatory variable X
- Simple linear regression is a good starting point
  - Slope estimate is likely biased due to omitted variables that effect both X and Y
- Experiments (RCTs) are ideal for determining causal relationship between X and Y
  - Costly and sometimes unfeasable
- Multiple regression can control for several relevant variables
  - Obtain causal relationship under conditional independance

$\hookrightarrow \varepsilon \perp X_1 \mid X_2, \ldots, X_k$