14.32 Recitation 3

Nina Wang

MIT Department of Economics

Fall 2023

Table of Contents

Table of Contents

Regression Basics: Terminology

$$Y_i = \alpha + \beta X_i + e_i$$

- α and β
  - **Regression constant** and **regression coefficient**
  - There can be as many coefficients as you want ($\beta_1, \beta_2, \beta_3$, etc.)
  - Sometimes, constant is denoted as β_0
  - The regression coefficient can be interpreted as the number of units Y changes by when X increases by one unit
  - The constant can be interpreted as the value of Y when $X = 0$ (sometimes doesn't make sense; we rarely interpret the constant).

$$Y_i = \beta_0 + \beta_1 X_i + e_i \text{ is equivalent to } Y_i = \alpha + \beta X_i + e_i$$

Regression Basics: Terminology

$$Y_i = \alpha + \beta X_i + e_i$$

- $\hat{Y}$: Predicted Value
  - We can think of the predicted value as the value of Y that our linear regression will predict for any given value of X.
  - $\hat{Y} = \alpha + \beta X_i$
  - $Y_i = \hat{Y} + e_i$

Regression Basics: Terminology

$$Y_i = \alpha + \beta X_i + e_i$$

- e_i: Residual
  - The difference between the observed value (Y_i) and the predicted value ($\hat{Y}$)
  - You can also think about the residual as the "noise" in the data or "error" of the model.
  - Sometimes denoted as ϵ_i
- Properties of the residual
  - $\Sigma e_i = 0 \rightarrow E[e_i] = 0$
  - $\Sigma X_i e_i = 0 \rightarrow E[X_i e_i] = 0$

Regression Basics: OLS Estimators

$$Y_i = \alpha + \beta X_i + e_i$$

- An OLS (Ordinary Least Squares) Regression will **minimize** the sum of the squared residuals.

$$\underset{\alpha,\beta}{\text{argmin}}\ \Sigma_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2$$

Regression Basics: OLS Estimators

$$Y_i = \alpha + \beta X_i + e_i$$

- An OLS (Ordinary Least Squares) Regression will **minimize** the sum of the squared residuals.

$$\Sigma e_i = 0$$

$$\underset{\alpha,\beta}{\text{argmin}}\ \Sigma_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2 \qquad \Sigma X_i e_i = 0$$

- By solving for the FOC for α and β, we get the following result:

$$\alpha = \bar{Y} - \beta\bar{X} \qquad \beta = \frac{Cov(X,Y)}{Var(X)}$$

Table of Contents

Determining Significance

- Your coefficient β can vary in magnitude, but usually what we're concerned about is whether or not this coefficient is **statistically significant.**
- We use the t statistic to measure statistic significance.

$$t = \frac{\hat{\beta} - \beta}{s.e(\hat{\beta})} \qquad s.e(\hat{\beta}) = \frac{\hat{\sigma}_{\beta}}{\sqrt{n}}$$

Determining Significance

- Your coefficient β can vary in magnitude, but usually what we're concerned about is whether or not this coefficient is **statistically significant.**
- We use the t statistic to measure statistic significance.

$$t = \frac{\hat{\beta}-\beta}{s.e(\hat{\beta})} \qquad\qquad s.e(\hat{\beta}) = \frac{\hat{\sigma}_\beta}{\sqrt{n}}$$

- Through our assumptions, we have that: $\frac{\hat{\beta}-\beta}{s.e(\hat{\beta})} \to N(0,1)$
- That means we can use the values of the Normal CDF to construct confidence sets and test for significance.

Determining Significance

- When testing for significance of a coefficient, our hypotheses are

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

- Thus, our t-statistic is $t = \frac{\hat{\beta}-0}{s.e(\hat{\beta})}$
- If $|t| > 1.96$, we reject the null. Otherwise, we fail to reject.

Example: reading Stata output

```
. reg weightloss health if sex == 0 // females only

      Source |       SS           df       MS      Number of obs   =       173
-------------+----------------------------------   F(1, 171)       =      6.30
       Model |  30.9558765         1  30.9558765   Prob > F        =    0.0130
    Residual |  839.694232       171  4.91049259   R-squared       =    0.0356
-------------+----------------------------------   Adj R-squared   =    0.0299
       Total |  870.650109       172  5.06191924   Root MSE        =     2.216

------------------------------------------------------------------------------
  weightloss | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      health |   .2327976   .0927192     2.51   0.013     .0497761    .4158191
       _cons |   3.374027   .1756569    19.21   0.000     3.027292    3.720763
------------------------------------------------------------------------------
```

Example: reading a table

Treatment effect on crude rate of prescription overdose

VARIABLES	(1) Crude Rate	(2) Crude Rate	(3) Crude Rate
treatment	1.106***	3.174***	-0.323
	(0.325)	(0.260)	(0.274)
Constant	3.900***	10.33***	6.788***
	(0.543)	(0.712)	(0.611)
Observations	690	740	690
R-squared	0.287	0.621	0.806
State FE	NO	YES	YES
Year FE	YES	NO	YES

Standard errors in parentheses

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: Center for Behavioral Health Statistics and Quality, SAMHSA,2011.

Constructing Confidence Intervals

- We also use the standard error to construct confidence intervals.
- For a 95% confidence interval (what we usually use), the confidence interval for β will be

$$[\beta - 1.96 \cdot s.e.(\beta), \beta + 1.96 \cdot s.e.(\beta)]$$

Judging our Model

- SSR (Sum of the Squared Residuals) Σe_i^2
- ESS (Explained Sum of Squares) $\Sigma(\hat{Y}_i - \bar{Y})^2$
- TSS (Total Sum of Squares) $\Sigma(Y_i - \bar{Y})^2$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

- the R^2 of a model measures the fraction of the variation of Y explained by the regression.

Table of Contents

Omitted Variable Bias

- Omitted variable bias occurs when there is a variable Z_i such that the two following conditions hold.
    - Z_i affects the outcome Y_i
    - Z_i is correlated with another regressor X_i

Omitted Variable Bias

- Direction of OVB (B is omitted variable, A is other regressor):

	A and B are positively correlated	A and B are negatively correlated
B is positively correlated with Y	Positive Bias	Negative Bias
B is negatively correlated with Y	Negative Bias	Positive Bias

Omitted Variable Bias: Example

Consider a regression of log wages (Y_i) on schooling (S_i), controlling for ability (A_i). This yields the following regression equation:

$$Y_i = \alpha + \rho S_i + \gamma A_i + e_i \text{ (long)}$$

Omitted Variable Bias: Example

Consider a regression of log wages (Y_i) on schooling (S_i), controlling for ability (A_i). This yields the following regression equation:

$$Y_i = \alpha + \rho S_i + \gamma A_i + e_i \text{ (long)}$$

Unfortunately, we're unable to observe ability (A_i). Thus, we have to make do with a regression on schooling alone.

$$Y_i = \alpha^* + \rho^* S_i + e_i^* \text{ (short)}$$

Omitted Variable Bias: Example

Consider a regression of log wages (Y_i) on schooling (S_i), controlling for ability (A_i). This yields the following regression equation:

$$Y_i = \alpha + \rho S_i + \gamma A_i + e_i \text{ (long)}$$

Unfortunately, we're unable to observe ability (A_i). Thus, we have to make do with a regression on schooling alone.

$$Y_i = \alpha^* + \rho^* S_i + e_i^* \text{ (short)}$$

- How do we find the value of the OVB?

Omitted Variable Bias: Example

Remember our two regressions:

$$Y_i = \alpha + \rho S_i + \gamma A_i + e_i \text{ (long)} \qquad Y_i = \alpha^* + \rho^* S_i + e_i^* \text{ (short)}$$

- OVB formula: $\rho^* = \rho + \underbrace{\gamma \delta_{AS}}$ ← OVB
  - δ_{AS} is coefficient of a regression of omitted (A_i) on included (S_i).

$A_i = \alpha + \beta S_i + e_i$

Omitted Variable Bias: Example

Remember our two regressions:

no bias / biased

$$Y_i = \alpha + \rho S_i + \gamma A_i + e_i \text{ (long)} \qquad Y_i = \alpha^* + \rho^* S_i + e_i^* \text{ (short)}$$

- OVB formula: $\rho^* = \rho + \gamma \delta_{AS}$
  - δ_{AS} is coefficient of a regression of omitted (A_i) on included (S_i).
- Another way of writing:
  - $OVB = \frac{Cov(S_i, e_i^*)}{Var(S_i)}$

$e_i^* = \alpha + \beta S_i + e_i$

Table of Contents

Proof of OLS coefficients Matrix OLS formula

$$\operatorname{argmin} \sum_i (Y_i - \alpha - \beta X_i)^2$$

FOC of α

$$\frac{\partial}{\partial\alpha} = -2\sum(Y_i - \alpha - \beta X_i) = 0$$

$$\sum(Y_i - \alpha - \beta X_i) = 0 \qquad \sum e_i = 0$$

$$\sum Y_i - n\alpha - \beta\sum X_i = 0$$

$$\alpha = \frac{1}{n}\sum Y_i - \beta\frac{1}{n}\sum X_i = 0$$

$$\alpha = \bar{Y} - \beta\bar{X}$$

$$\frac{\partial}{\partial\beta} = -2\sum(Y_i - \alpha - \beta X_i)X_i = 0$$

$$\sum Y_i X_i - \alpha\sum X_i - \beta\sum X_i^2 = 0$$

$$\sum Y_i X_i - (\bar{Y} - \beta\bar{X})\sum X_i - \beta\sum X_i^2 = 0$$

$$" \quad - \frac{\sum Y_i}{n}\sum X_i + \beta\frac{\sum X_i}{n}\sum X_i \quad "$$

$$\sum Y_i X_i - \frac{\sum Y_i}{n}\sum X_i - \beta\left(\sum X_i^2 - \frac{\sum X_i}{n}\sum X_i\right) = 0$$

$$\beta = \frac{\sum Y_i X_i - \frac{\sum Y_i}{n}\sum X_i}{\sum X_i^2 - \frac{\sum X_i}{n}\sum X_i}$$

$$= \frac{\frac{1}{n}\sum Y_i X_i - \frac{1}{n}\sum Y_i \frac{1}{n}\sum X_i}{\frac{1}{n}\sum X_i^2 - \frac{1}{n}\sum X_i \frac{1}{n}\sum X_i}$$

$$\frac{1}{n}\sum Y_i = \bar{Y} = E[Y]$$

$$= \frac{E[Y_i X_i] - E[Y_i]E[X_i]}{E[X^2] - (E[X])^2} = \frac{Cov(X,Y)}{Var(X)}$$

Practice Problem

(c) Consider the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$. Transform the regression so that you can use a t-statistic to test

i $H_0 : \beta_1 = \beta_2$

$\beta_1 - \beta_2 = 0 \quad (\beta_1 - \beta_2)(X_{1i} - X_{2i})$

$$\left.\frac{\beta}{s.e.(\beta)}\right]$$

$b_1 \quad b_2 \Big] - \beta_1 - \beta_2$

$Y_i = \alpha + b_1 Z_i + b_2 Z_2 + e_i$

$\beta_1 = b_1 + b_2$

$\beta_2 = b_1 - b_2$

$$b_2 = \frac{\beta_1 - \beta_2}{2}$$