Stata Training Camp

Lecture 3 Regression Analysis and Output

Juncheng Jiang

jc_jiang@stu.zuel.edu.cn

School of Public Finance and Taxation, Zhongnan University of Economics & Law
Associate Editor, Student Academic Research Center

2024.12.01

Review: Data Classification

Generally speaking, economic data can be divided into three/four categories:

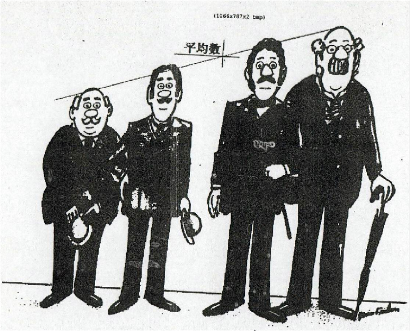
- Cross-section data
  - Consists of a sample of individuals, households, firms, cities, countries, or various other units taken at a given point in time.
  - Minor time differences are ignored, e.g., surveying different people in different months within a year.
- Time series data
  - Consists of observations on one or more variables over time.
  - It is rarely assumed that economic data are independent across time.
  - Need to pay attention to the data frequency at the time of collection (Daily/Weekly/Monthly/Quarterly/Annually?).
- Pooled cross section
  - Has both cross-sectional and time-series characteristics. E.g., A group of people was surveyed in 2022, and a new group of people was surveyed in 2024.
- Panel data
  - Consists of a time series for each cross-sectional unit in the dataset.

Regression: regression

Classical Meaning of Regression:
Galton's concept of regression in genetics (relationship between parents' height and children's height). Children's height tends to "regress" towards the average human height.

Modern Meaning of Regression:
A study of the dependence of a dependent variable on several explanatory variables.

Purpose (Essence) of Regression:
Estimating the average value of the dependent variable from the explanatory variables.



OLS: Ordinary Least Squares

Taking simple linear regression as an example:

$$y = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, 2, 3 \ldots n)$$

- y_i: dependent variable, regressand
- x_i: explanatory variable, independent variable, regressor
- α: intercept term or constant term;
- β: slope
- α and β: collectively called "regression coefficients" or "parameters".
- ε_i: "error term" or "disturbance term", including omitted other factors, measurement errors in variables, specification errors in the regression function (e.g., neglecting nonlinear terms), and the inherent randomness of human behavior, etc.

OLS: Principle of Ordinary Least Squares

$$y_i = \beta_0 + \beta_1 x_{1_i} + \varepsilon_i$$

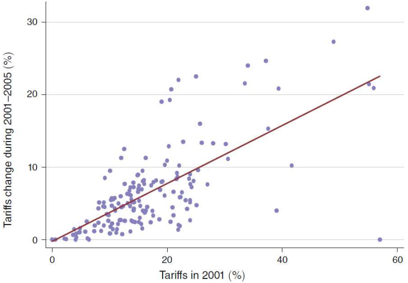
Minimizing the sum of squared residuals:
$\min \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1_i})^2$

The sample regression function is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1_i}$$

i.e.

$$y_i = \hat{y}_i + \hat{\varepsilon}_i$$



Derivation of R^2

Analyze the relationship between the observed values Y_i, the fitted values $\hat{Y}_i$, and the mean $\overline{Y}$:

$$Y_i - \bar{Y} = (Y_i - \bar{Y}) + \hat{Y}_i - \hat{Y}_i = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

Squaring both sides of the equation and summing, it can be proven (Hint: cross term $\sum(\hat{Y}_i - \bar{Y})e_i = 0$):

$$\underset{\text{(TSS)}}{\sum}(Y_i - \overline{Y})^2 = \underset{\text{(ESS)}}{\sum}(\hat{Y}_i - \overline{Y})^2 + \underset{\text{(RSS)}}{\sum}(Y_i - \hat{Y}_i)^2$$

Or expressed as $\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2$

- Total Sum of Squares $\sum y_i^2$ (TSS): The sum of squared deviations of the observed values of the dependent variable Y from their mean (total sum of squares) (Indicates the total degree of variation in Y).
- Explained Sum of Squares $\sum \hat{y}_i^2$ (ESS): The sum of squared deviations of the fitted values of the dependent variable Y from their mean (regression sum of squares).
- Residual Sum of Squares $\sum e_i^2$ (RSS): The sum of squared

What is R^2?

What is R-squared?

Definition: The proportion of the regression sum of squares (ESS) $\sum \hat{y}_i^2$ in the total sum of squares (TSS) $\sum y_i^2$ is called the coefficient of determination, denoted by r^2 or R^2:

$$R^2 = \frac{\sum \hat{y}^2}{\sum y_i^2} \quad \text{or} \quad R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

Its value ranges from 0 to 1. It is a non-negative statistic, also known as: goodness of fit. It indicates how well the model fits the sample observations.

However, R^2 is not particularly important and is generally not reported in articles qwq.
What's more important are the t-value and p-value.

Standard Errors of Regression Coefficients

In a regression model, the standard deviation of the estimated regression coefficient is generally called the standard error. The standard error of a regression coefficient is used to measure the precision and stability of the regression coefficient estimate. The smaller the standard error, the higher the overall precision of the regression coefficient. In other words, it is the standard deviation of the sampling distribution of the regression coefficient.

Regression

There is a relationship you need to know when reporting regression analysis results and conclusions.

$$\frac{\text{Regression Coefficient}}{\text{Standard Error}} = t\text{-value}$$

Based on the t-value or t-distribution, the significance P-value for the influence of a specific independent variable on Y can be obtained.

Therefore, a common format we see in papers is: coefficient (se). Using the standard error of the regression coefficient, a confidence interval for the coefficient can be constructed. The confidence interval provides a range of values within which the true value is considered statistically plausible.

Method for Testing Regression Coefficients: z-test

Establish Hypotheses: The null hypothesis is

$$H_0 : \beta_2 = 0$$

The alternative hypothesis is

$$H_1 : \beta_2 \neq 0$$

(Essence: Test whether $\beta_2 = 0$, i.e., test whether X_i has a significant effect on Y)

—

(1) When σ^2 is known or the sample size is large enough The normal distribution can be used for the z-test:

$$z^* = \frac{\hat{\beta}_2 - \beta_2}{\mathrm{SE}(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{\mathrm{SE}(\hat{\beta}_2)} \sim N(0,1)$$

Given α, look up the critical value z in the normal distribution table.

- If $-z < z^* < z$, do not reject the null hypothesis H_0.
- If $z^* < -z$ or $z^* > z$, reject the null hypothesis H_0.

Method for Testing Regression Coefficients: t-test

(2) When σ^2 is unknown and the sample size is small
σ^2 must be replaced by $\hat{\sigma}^2$, and the t-distribution can be used for the t-test:

$$t^* = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\mathrm{SE}(\hat{\beta}_2)}} = \frac{\hat{\beta}_2}{\mathrm{SE}(\hat{\beta}_2)} \sim t(n-2)$$

Given α, look up $t_{\alpha/2}(n-2)$ in the t-distribution table.

- If $t^* \leq -t_{\alpha/2}(n-2)$ or $t^* \geq t_{\alpha/2}(n-2)$,
  then reject the null hypothesis $H_0 : \beta_2 = 0$ and do not reject the alternative hypothesis $H_1 : \beta_2 \neq 0$.
- If $-t_{\alpha/2}(n-2) \leq t^* \leq t_{\alpha/2}(n-2)$,
  then do not reject the null hypothesis $H_0 : \beta_2 = 0$.

P-value

Method for Determining Parameter Significance Using the *P*-value

Procedure: Compare the given significance level α with the *p*-value:

- If $\alpha >$ *p*-value, then it must be that $|t^*| > |t_{\alpha/2}|$. In this case, reject the null hypothesis $H_0 : \beta_k = 0$ at the significance level α. This suggests that X has a significant effect on Y.
- If $\alpha \leq$ *p*-value, then it must be that $|t^*| \leq |t_{\alpha/2}|$. In this case, do not reject the null hypothesis $H_0 : \beta_k = 0$ at the significance level α. This suggests that X does not have a significant effect on Y.

Rule: When $p < \alpha$, the *p*-value is smaller, and it indicates stronger evidence to reject the null hypothesis H_0.

Interpretation of Coefficients in Different Forms

For any basic linear regression, sometimes the raw values are used directly in the regression, and sometimes they are log-transformed. Below are common (log Œ linear) combinations. The interpretation of β_1 differs across these forms.

$$y = \beta_1 x + \beta_0 + \varepsilon$$

$$y = \beta_1 lnx + \beta_0 + \varepsilon$$

$$lny = \beta_1 x + \beta_0 + \varepsilon$$

$$lny = \beta_1 lnx + \beta_0 + \varepsilon$$

β_1 indicates:

X	Y
x (unit value)	y (β_1 unit value)
lnx (100%)	y (β_1 unit value)
x (unit value)	lny ($100\beta_1$%)
lnx (1%)	lny (β_1%)

Fixed Effects FE

First, fixed effects are control variables that can influence both the explanatory variables and the dependent variable.

Second, fixed effects are influencing factors (control variables) with constant, unchanging characteristics, including both observable and unobservable ones.

Finally, fixed effects have two dimensions: one is space (entity), and the other is time.

Fixed Effects

For a regression model:

$$y_{it} = \beta_0 + \beta_1 x_{1_{it}} + \beta_2 x_{2_{it}} + \gamma_i + \delta_t + \varepsilon_{it}$$

Averaging i respect to t:

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{1_i} + \beta_2 \bar{x}_{2_i} + \gamma_i + \bar{\varepsilon}_i$$

Thus

$$\gamma_i = \bar{y}_i - \beta_0 - \beta_1 \bar{x}_{1_i} - \beta_2 \bar{x}_{2_i} - \bar{\varepsilon}_i$$

Rearranging the original model:

$$\ddot{y}_{it} \equiv y_{it} - \bar{y}_i = \beta_0 + \beta_1 \ddot{x}_{1_{it}} + \beta_2 \ddot{x}_{2_{it}} + \delta_t + \ddot{\varepsilon}_{it}$$

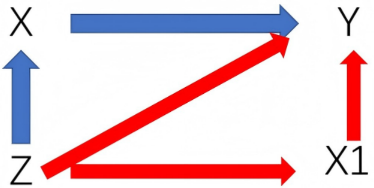
That's "How to control the individual FE", and the same to time FE.

Fixed Effects

Several types: individual fixed effects, time fixed effects, province fixed effects, industry fixed effects, and interacted fixed effects, etc.

Grouping is based on unique, time-invariant characteristics. For example, individual fixed effects represent the influence of each individual's unique, time-invariant characteristics on the dependent variable. Province fixed effects represent the influence of each province's unique, time-invariant characteristics on the dependent variable. Different groupings can form different fixed effects.

Instrumental Variables: A Method to Address Endogeneity



Requirements for an Instrumental Variable (Z): It can only affect Y through X

- Z can affect X (Z medicine is effective for X disease; cannot be flour or plain water)
- Z cannot affect Y through other pathways
  - Z cannot directly affect Y (If Z has side effects like insomnia or drowsiness, it is not a good instrument)
  - Z cannot affect Y through other variables (If Z causes headaches, it is not a good instrument)

2SLS

Two-Stage Least Squares (2SLS)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

x_2 is endogenous, and the instrumental variable is z

- Applicable condition: No autocorrelation or heteroskedasticity in the error term
- First stage: regress x_2 on x_1 and z, obtain $\hat{x}_2$
- Second stage: regress y on x_1 and $\hat{x}_2$
- Do not perform the two regressions manually yourself (this would lead to incorrect residuals)

Stata Commands:

- ivregress 2sls y x_1 $(x_2 = z_1 z_2)$
- ivreg2 y x_1 $(x_2 = z_1 z_2)$

What You Will Learn in This Lecture:

```
reg
ivreg
xtreg
reghdfe
ivreghdfe
est // est store, est tab, etc.
outreg2
.......
```

Some More Comprehensive References (Wechat Link)

- Fixed Effects: "Fixed Effects: The clearest post explaining it so far, lifesaver!" https://mp.weixin.qq.com/s/5hxLz7XJ30Z9Jmo9SGEWYA
- "Endogeneity Topic | Instrumental Variables and 2SLS — ivregress2 Command" https://mp.weixin.qq.com/s/HCNdnkO0P8Wb0cHnGINWZg
-
- "Stata: FAQ on the reghdfe Command" https://mp.weixin.qq.com/s/fFLCRphoJSgQEMGEPpqfWw
- "Stata Output: Detailed Explanation of the outreg2 Command" https://mp.weixin.qq.com/s/jcnMO2IMiqsDg0SRcKzEMg

Launch Stata

DO!