Regression analysis

Lars Nesheim

Centre for Microdata Methods and Practice (CeMMAP), UCL

November 2011

Introduction

- Highly **incomplete and fragmented** introduction to regression analysis.
- Have **data** on y and x (x is high dimensional vector) and would like to:
  - Study **correlation** between y and elements of x.
  - Use x to **forecast, predict or impute** y.
  - Understand how much of **variation** in y is "explained" by x.
  - Estimate **causal impact** of x on y after controlling for confounding factors

Introduction (2)

- Elements of X.
  - Can be **continuous or discrete.**
  - May be **measured with error** (this causes problems).
- Elements of Y.
  - Can be continuous.
  - Can be discrete.
    - Usually requires a **non-linear** model.
    - However, researchers often use the linear probability model.
    - It is much better to use a discrete outcome model.
  - May be measured with error (this is less of a problem).
  - May be censored or truncated.
    - Usually requires use of **non-linear** model.

Main methods

1. **Parametric methods**
    1. Ordinary least squares
    2. Maximum likelihood
    3. Method of moments
    4. Quantile regression
    5. Bayesian methods
2. **Semi- and Non-parametric methods**
    - Allow "parameters" to be infinite dimensional.
    - Estimate $E[Y|X] = f(X)$ to be an unknown function instead of a known one like $x\beta$.
    - Estimate probability density function of x rather than estimate mean and variance.

Basic linear model (notation)

- **Data** on (y_i, x_i) for $i = 1, ..., N$ where $y_i \in \mathbf{R}$ and $x_i \in \mathbf{R}^k$.
- Let $Y = (y_1, ..., y_N)^T$ be a $(N \times 1)$ vector of **outcomes** and let X be an $(N \times K)$ matrix of **regressors**. That is,

$$X = \begin{bmatrix} x_1(1) & \cdots & x_1(K) \\ \vdots & & \vdots \\ x_N(1) & \cdots & x_N(K) \end{bmatrix}.$$

- Let $\varepsilon = (\varepsilon_1, ..., \varepsilon_N)$ be an $(N \times 1)$ vector of **errors** or **unobserved** variables.

Basic linear model

- The **linear model** is
$$Y = X\beta + \varepsilon$$
where β is a $(K \times 1)$ vector of **parameters** to be estimated.
  - Usually the model includes a constant so that $x_i(1) = 1$ for all i.
  - Key restriction is that the model is linear in β (can allow for example x and x^2 or $\log(x)$.)
  - X may include "dummy" variables that indicate membership in a group. For, example $x_i(2) = 1$ if i is female and $x_i(2) = 0$ otherwise.

Basic linear model (goals)

- **Goals:**
  - **Unbiased** or **consistent** estimates of β.
  - **Prediction** of Y.
  - **Analysis of variance** of Y.
  - **Test of hypotheses** about β.

Basic linear model (problems)

1. **Mis-specification.** Suppose the correct model is not linear?
   1. Non-linear models or discrete outcomes.
   2. Censoring or truncation of of outcomes
2. **Endogeneity.** What if X is correlated with ε?
   1. Omitted variables.
   2. Measurement error.
   3. Joint causation.
3. **High dimensional data.** What can one do if K is large?
4. **Robustness.** How to reduce influence of outliers in data?
5. **Correlation in errors.** How do you correct for correlation in errors?
6. **Non-random sample.** How does one weight the data?

Estimation in linear models

- **Ordinary least squares**. Choose β to solve the least squares problem

$$\min_{\{\beta\}} \left\{ 0.5 \left(Y - X\beta \right)^T \left(Y - X\beta \right) \right\}$$

- **First order conditions** are

$$X^T X\beta - X^T Y = 0.$$

- **Estimator** of β is

$$\widehat{\beta} = \left(X^T X \right)^{-1} X^T Y$$

  - Requires that $X^T X$ has rank K.

Properties of estimator (unbiased)

- **Assume** that
$$E\left[\varepsilon \,|X\right] = 0.$$
- **Then**
$$\begin{aligned}
E\left[\widehat{\beta}\,|X\right] &= E\left[\left(X^T X\right)^{-1} X^T Y\,|X\right] \\
&= E\left[\left(X^T X\right)^{-1} X^T \left(X\beta + \varepsilon\right)|X\right] \\
&= E\left[\left(X^T X\right)^{-1} \left(X^T X\right)|X\right]\beta + E\left[\varepsilon\,|X\right] \\
&= \beta.
\end{aligned}$$

Asymptotic normality

- Further, **assume** that
$$V\left(\varepsilon\,|X\right)=\sigma^2 I.$$
- **Then**
$$V\left(\widehat{\beta}\right)=\widehat{\sigma}^2\left(X^TX\right)^{-1}$$
where
$$\widehat{\sigma}^2=\frac{1}{n}\widehat{e}^T\widehat{e}$$
is estimate of variance of error.and where
$$\widehat{e}=Y-X\widehat{\beta}.$$
- In a **large sample**, (under very general conditions), the **central limit theorem** can be used to show that
$$\sqrt{N}\left(\widehat{\beta}-\beta\right)\xrightarrow{A}N\left(0,\widehat{\Sigma}\right).$$

Confidence intervals and hypothesis tests

- These results can be used to construct confidence intervals for β and to conduct hypothesis tests.
- When β is a scalar, a **95% confidence interval** for β is

$$\widehat{\beta} \pm 1.96\widehat{\sigma}_\beta.$$

- **Test the hypothesis** that $\beta_1 + \beta_2 = 0$.
    1. Let $s = \beta_1 + \beta_2$.
    2. Then $\widehat{s} = \widehat{\beta}_1 + \widehat{\beta}_2$ converges in distribution to a normal random variable with mean $s = \beta_1 + \beta_2$ and variance $\sigma_s^2 = \sigma_{11} + \sigma_{22} + 2\sigma_{12}$.
    3. **Reject the hypothesis** if $\frac{\widehat{s}}{\widehat{\sigma}_s} \geq 1.96$.

Prediction

- **Best predictor** of Y is

$$E\left[Y\,|X\right] = X\widehat{\beta}.$$

- Minimises the sum of squared prediction errors.

Goodness of fit

- **Coefficient of determination** or $\mathbf{R}^2$ measures the fraction of variance of the outcome that is explained by the model. It is

$$\mathbf{R}^2 = 1 - \frac{e^T e}{(y - \overline{y})^T (y - \overline{y})}.$$

- A measure of "**Goodness-of-fit**".
- When $\mathbf{R}^2$ is near zero, then most of variance is explained by errors.
- When near one, most of variance is explained by model.
- When variables are added to model, $\mathbf{R}^2$ increases. So, researchers often used "adjusted $\mathbf{R}^2$

$$\overline{\mathbf{R}}^2 = 1 - \left(1 - \mathbf{R}^2\right) \frac{n-1}{n-k-2}$$

which adjusts $\mathbf{R}^2$ for the number of variables in the model.

Other topics

1. Maximum likelihood estimation.
   - Linear models.
   - Nonlinear models.
   - Discrete outcomes.
2. Instrumental variables methods.
3. Systems of equations.
4. Penalized methods.