PoliSci 4782 Political Analysis II

The Overview of Statistical Models

Seamus Wagner

The Ohio State University

Notation

- **Outcome variable**
  - Y is a variable in the abstract ($n \times 1$ in our sample data)
  - y_i is a realized value of this variable (after we observe it)
  - Y_i is an unobserved value of this variable (whose value is still random before we actually observe it)
- **Explanatory variables**
  - $\boldsymbol{X}$ is the whole set of our explanatory variables ($n \times k$ in our sample data, n is the number of observations, k is the number of variables)
  - $x_{i,j}$ is a realized value of variable j in observation i
  - In statistical analysis, $\boldsymbol{X}$ is fixed, not random (because we already collect data and feed them to our models)

Linear Regression Notation

- **The basic expression**:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_k X_{i,k} + \epsilon_i$$

- **A more succinct expression**:

$$Y_i = \beta \boldsymbol{X_i} + \epsilon_i$$

ϵ_i is regression residual, capturing the difference between the actual Y_i and the predicted $\hat{Y}_i$ by $\beta \boldsymbol{X_i}$.

Two Components of Linear Regression

- **The system component**:

$$\beta \boldsymbol{X_i} = \hat{Y}_i$$

- **The stochastic component**:

$$\epsilon_i \sim N(0, \sigma^2)$$

On average, the expected Y_i should equal to the predicted $\hat{Y}_i$, so the average residual should be 0, with a certain amount of stochastic errors captured by σ^2 in each case.

Alternative Notation

- **The system component**:

$$\mu_i = \beta \boldsymbol{X_i}$$

- **The stochastic component**:

$$Y_i \sim N(\mu_i, \sigma^2)$$

On average, the expected Y_i should equal to the predicted $\hat{Y}_i$ whose value is written as μ_i; the actual value of Y_i follows a normal distribution centered at μ_i with the variance of σ^2.

Systematic and Stochastic Components

$\mu_i = \beta \boldsymbol{X_i}$ and $Y_i \sim N(\mu_i, \sigma^2)$:

Generalized Model Notation

The system component: $\mu_i = g(X_i, \beta)$
The stochastic component: $Y_i \sim f(\mu_i, \eta)$

- μ_i is a systematic feature of the probability density of Y_i (the mean of Y_i in linear regression)
- β is effect parameter (coefficients on variable X_i)
- $g(\cdot)$ is a linear or *nonlinear* function to put together variables and effect parameters
- $f(\cdot)$ is a probability distribution that is not necessarily normal (binomial, Poisson, etc.)
- η is ancillary parameter (a constant feature of the probability density f across i, which governs the shape of the distribution)

Varieties of Systematic Components

- $\mu_i = g(X, \beta) = \beta_0 + \beta_1 X$, $g(X, \beta)$ is linear
- In linear regression, $E(Y) = \mu = \beta_0 + \beta_1 X$

Varieties of Systematic Components

- $\mu_i = g(X, \beta) = \beta_0 + \beta_1 X + \beta_2 X^2$, $g(X, \beta)$ is still linear
- In linear (quadratic) regression, $E(Y) = \mu = \beta_0 + \beta_1 X + \beta_2 X^2$

Varieties of Systematic Components

- $\mu_i = g(X, \beta) = \frac{1}{1+e^{-X\beta}}$, $g(X, \beta)$ is *nonlinear*
- It is called logistic regression, used to model the probability of a binary outcome variable, $Prob(Y = 1) = \mu = \frac{1}{1+e^{-X\beta}}$

Model Specification/Choosing a “Right” Function

- Be informed by theory: what do your domain knowledge and literature say about the outcome variable?

- Understand your data: what does the distribution of your data in the outcome variable look like (exploring data with plots is always a good idea)?

- But a certain amount of specification error is common.

Specification Errors

- some specification errors, but not terribly wrong or bias
- still a not bad approximation of the truth in general
- but will certainly lead to wrong/unrealistic predictions for some x

Varieties of Stochastic Components

- $Y \sim N(\mu, \sigma^2)$, in theory the value of Y is unbounded and continuous
- This is what we use for linear regression

Varieties of Stochastic Components

- $Y \sim Binom(\pi, n)$, where $Prob(Y = 1) = \pi$ and n is the number of "trails"
- We use this to model binary outcome variables (discrete)

Varieties of Stochastic Components

- $Y \sim Poiss(\lambda)$, where $E(Y) = Var(Y) = \lambda$
- We use this to model count outcome variables (discrete)

Choosing a "Right" Stochastic Component

- Understanding your outcome variables (discrete or continuous, bounded or unbounded, etc.)
- Some rules of thumb that we will go through in coming sessions (logit/probit for binary outcomes, Poisson/negative binomial for count, etc.)
- You can design and customize your own model that fits with your data and theory
- But a certain amount of errors is inevitable.

Forms of Uncertainty

The system component: $\mu_i = g(X_i, \beta)$
The stochastic component: $Y_i \sim f(\mu_i, \eta)$

- **Estimation uncertainty**: uncertainty about the true values of β and η
    - it is indicated by estimated *standard errors* of those parameter in our models, which decreases with our sample size.

- **Fundamental uncertainty**: fundamental complexity and randomness of the reality, captured by η in the stochastic component
    - things happen not in a perfectly deterministic way, so fundamental uncertainty exists no matter what

What Comes Next?

- more detailed discussion about outcome variables, their probability distributions, and model specification
- estimates on effect parameters and estimation uncertainty measures (standard errors)
- refresher on linear regression (a special case of generalized linear regression)