IQSR Analyzing Survey Data (Part II)

Chao-Yo Cheng

Birkbeck
UNIVERSITY OF LONDON

Plan ahead

- Weeks 4 and 5: Analyzing survey data
  - Week 4: Survey design and descriptive analysis of survey data.
  - **Week 5: Inferential (regression) analysis of survey data**.
- Weeks 8 and 9: Multilevel models
- Week 10: Review for exams

Outline for today

- Week 5 takeway: ***"To use weights or not, that is the question"***
- Regression analysis of survey data
- Revisiting logistic (logit) regression – how we get from OLS to GLM
- Live session: Studying the correlates of public views towards abortion

Week 5 lecture takeway

- Survey is one of the most common social research techniques; researchers seek to draw useful information about the **population** based on a **sample**.
- A survey needs to be carefully designed to ensure sure the "sample" is **representative** of the population. **Simple random sampling** is a common sampling approach.
- We use **survey weights** to correct or adjust the unrepresentative sample so the results can speak to the population.
- Without weights, our findings can be **biased** (due to the unrepresentative sample) or can only apply to a unique subset of population (i.e., **not generalizable**).

Week 5 live session recap

- In `R`, we use functions provided by packages "`survey`" and "`srvyr`" so as to include weights in our analysis.
  - First, turn the data frame into a **survey** object.
  - Next, use the summary statistics functions (e.g., `survey_total()`) in both packages to carry out descriptive and inferential analysis.
- Check out the package manuals and vignettes:
  - `survey`: https://r-survey.r-forge.r-project.org/survey/.
  - `srvyr`: http://gdfe.co/srvyr/index.html.
- The UCLA handout (link on Moodle) has a work example for the use of `survey` package.

Regression analysis of survey data

- Varieties of model choices – depending on the response variables.

Regression analysis of survey data

- Varieties of model choices – depending on the response variables.
  - Continuous variable: OLS
  - Binary variable: Logit/probit
  - Ordinal variable: Ordered logit/probit
  - Categorical variable: Multinomial logit/probit

Regression analysis of survey data

- Varieties of model choices – depending on the response variables.
  - Continuous variable: OLS
  - Binary variable: Logit/probit
  - Ordinal variable: Ordered logit/probit
  - Categorical variable: Multinomial logit/probit
- Resources:
  - "R Data Analysis Examples" by UCLA Statistical Methods and Data Analytics (https://stats.oarc.ucla.edu/other/dae/). – you can find many practical examples here.
  - "Regression and Other Stories" (2020) by Aki Vehtari, Andrew Gelman, and Jennifer Hill. – if you need a comprehensive coverage.

Regression analysis of survey data

"Essentially, all models are wrong, …but some are useful."

- George Box

(One of the most influential statisticians of the 20th century and a pioneer in the areas of quality control, time series analysis, design of experiments and Bayesian inference.)

ORACLE

"All models are wrong;" be an informed quantitative researcher.

Birkbeck
UNIVERSITY OF LONDON

Logit regression revisited

- Steps for today's review (when you read through the slides, feel free to start from anywhere):
  - Step 1: OLS recap (3 slides).
  - Step 2: From OLS to GLMs (1 slide).
  - Step 3: Logit as the workhorse GLM (6 slides).
- Not covered here: Model assumptions, diagnostics, and comparison (they will be discussed in the review workshop).

Step 1: OLS recap

- "Linear regression" (or **ordinary least squares**, aka OLS) means we use a **linear** (i.e., straight line) function to model the relationship between X (predictors) and Y (dependent variable).
- A simple **bivariate** (i.e., two-variable or one X) OLS can be specified as a linear function

$$Y = \alpha + \beta X + \epsilon,$$

  where α and β are the **intercept** (or constant) and the **slope** (or the regression coefficient) of the linear function.
- OLS aims to find the **"best"** fit of the regression line based on our data.
  - In plain English, the estimated $\hat{\alpha}$ and $\hat{\beta}$ will be a line that produced the smallest residuals/errors.
  - Formally speaking, the estimated regression line will produce the smallest **sum of squared errors** or **sum of squared residuals** (this is where the name of OLS came from).

Step 1: OLS recap

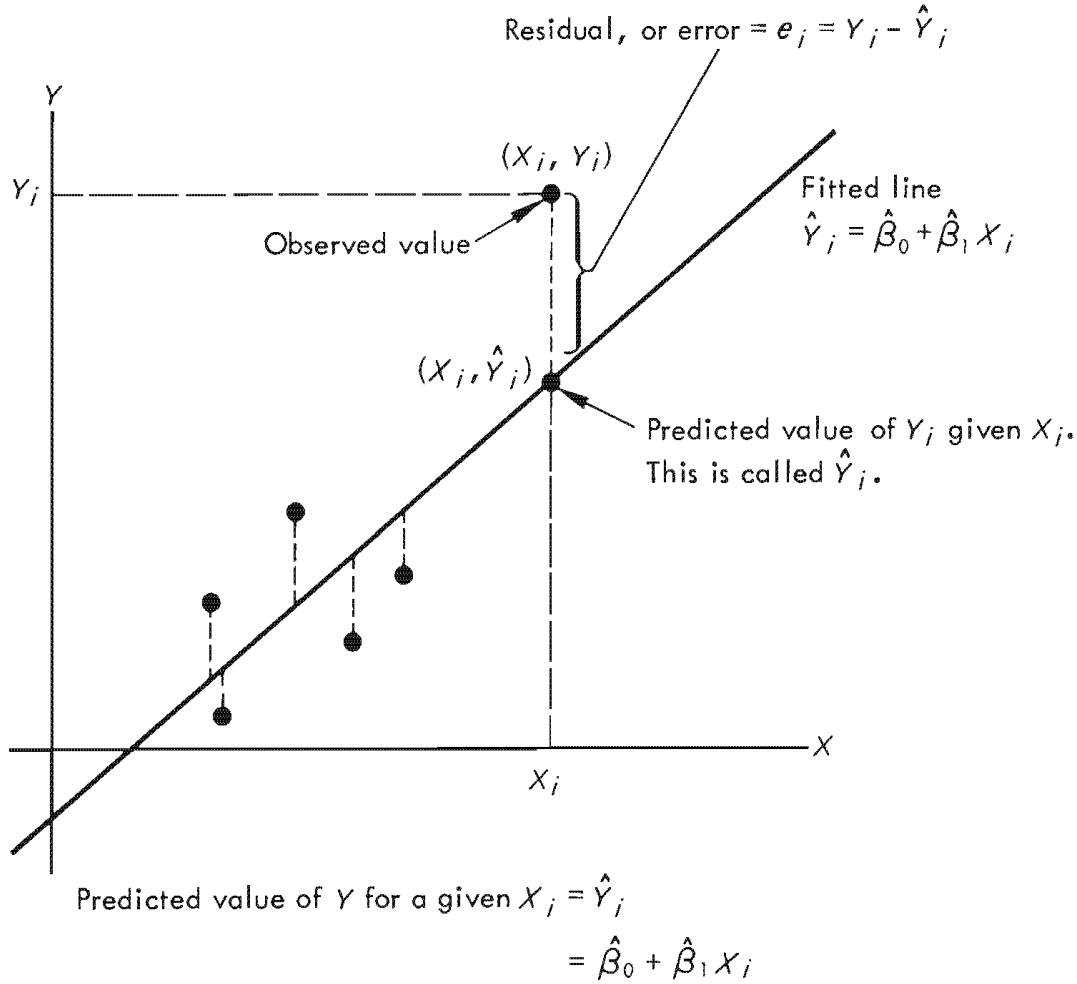


FIGURE 3-3 Notation for least-squares regression

Birkbeck
UNIVERSITY OF LONDON

Step 1: OLS recap

Consider the following linear regression model:

$$\text{Vote} = 50 + 2 \times \text{Hours}.$$

The model shows the relationship between **Hours** (the number of hours a party spends on campaigning during the election) and **Vote** (the percentage of votes won by the party).

- We can (attempt to) use this model to predict a party's vote share. For instance, when Hours$=0$, Vote$=50$; when Hours$=1$, Vote$=52$.
- A party will get 50% of votes without doing any campaigning at all.
- A party can increase its vote share 2 percentage points (i.e., from 50% to 52% of votes) if they spend one additional hour campaigning.

Step 2: From OLS to GLM

- **OLS will produce weird predictions** when Y is binary (i.e., each observation can only take one of the two values, say either 0 or 1) or bounded.
  - Using the same example, Vote= 110 with Hours= 30; no party can get 110% of the votes.
  - From the probability perspective, OLS may generate predictions larger than 1 and lower than 0 (and these predictions make no sense in real life).
- **Key OLS assumptions can be violated** when Y is binary, such as
  - **Normality**: OLS may create errors that are not normally distributed.
  - **Homoscedasticity**: OLS may create predictable errors (using X) rather than just random noise.

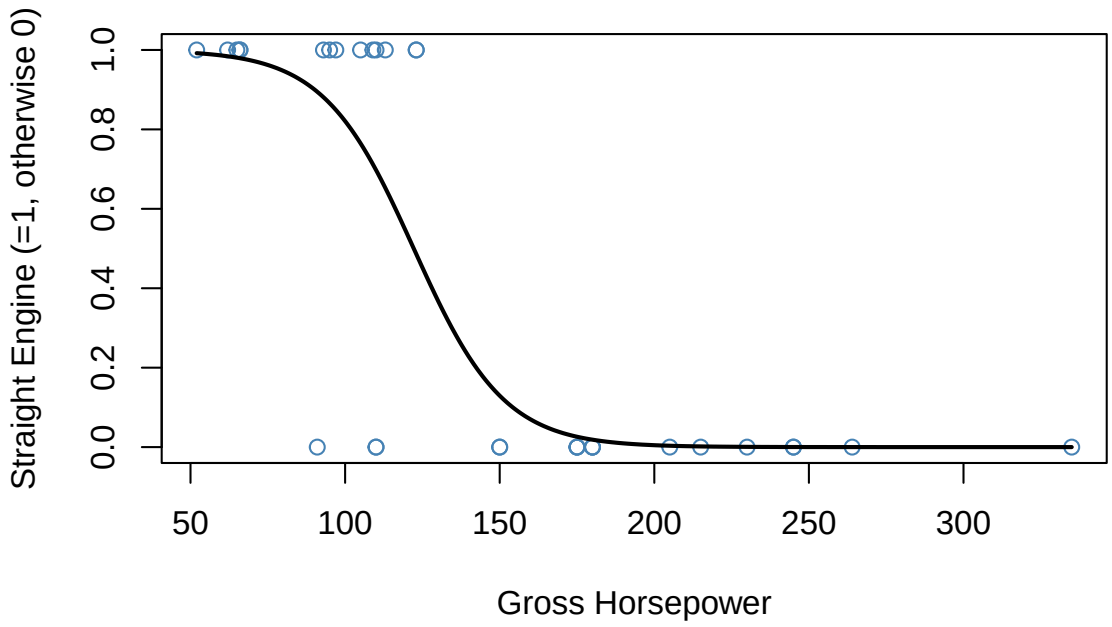
Step 2: From OLS to GLM

- **OLS will produce weird predictions** when Y is binary (i.e., each observation can only take one of the two values, say either 0 or 1) or bounded.
  - Using the same example, Vote= 110 with Hours= 30; no party can get 110% of the votes.
  - From the probability perspective, OLS may generate predictions larger than 1 and lower than 0 (and these predictions make no sense in real life).
- **Key OLS assumptions can be violated** when Y is binary, such as
  - **Normality**: OLS may create errors that are not normally distributed.
  - **Homoscedasticity**: OLS may create predictable errors (using X) rather than just random noise.
- Further ref: Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Los Angeles, CA: Sage – see pp. 38-40.

Step 3: Logit regression

- Logit regression is used to model a **binary** outcome or a probability, which we denote as p. By definition,
  - A probability p can only take the values between 0 and 1 (e.g., whether or not two countries fight each other) such that $p = 1$ means an event takes place.
  - The probability of something happening (e.g., two countries fight) and that of some not happening (e.g., two countries do not fight) will sum up to 1.
- Logit regression is a **generalized linear model** (GLM), meaning that we are using a **linear function** to model a **non-linear relationship** between X and Y (see an example on the next slide).

Step 3: Logit regression



The correlation between gross horsepower and whether or not a car has straight engine, using logit regression. Each blue dot is an observation in the dataset.

Birkbeck
UNIVERSITY OF LONDON

Step 3: Logit regression

- Logit regression uses the **logit** link function, which transform p (probability) into $\log\left(\frac{p}{1-p}\right)$ (the log-odds):

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

- A simple bivariate logit regression can be specified as follows:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta X + \epsilon.$$

- **Let's unpack the bivariate logit regression step by step.** We will start with the left hand side before moving on to the right hand side.

Step 3: Logit regression

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta X.$$

- If p is the probability of an event, then the odds is $\frac{p}{1-p}$, which suggests the chance of an event taking place relative to the opposite scenario.
  - Say today's probability of raining is 0.8, $p = 0.8$ (the **probability** of raining) and $1 - p = 0.2$ (the probability of not raining).
  - Next, the **odds** of raining is then $\frac{p}{1-p} = \frac{0.8}{1-0.2} = 4$.
  - Finally, the **log of odds** here is thus $\log\left(\frac{p}{1-p}\right) = \log(4)$ or the log of 4 with e (a mathematical constant) as the base.
- **Logit link function allows us to transform p, the original dependent variable, into log-odds such that the dependent variable is no longer bounded between 0 and 1**; it can also go negative.

Step 3: Logit regression

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta X \quad \text{a.k.a} \quad \log(\text{Odds}) = \alpha + \beta X$$

One-unit increase in X (e.g., moving X from 0 to 1) corresponds to

- β changes in $\log\left(\frac{p}{1-p}\right)$, or the **log odds**, such that

$$\begin{aligned} \beta &= \log(\text{Odds When X=1}) - \log(\text{Odds When X=0}) \\ &= \log\left(\frac{\text{Odds When X=1}}{\text{Odds When X=0}}\right). \end{aligned} \tag{1}$$

- The exponent of β will return the odds-ratio (OR), or

$$e^{\beta} = \frac{\text{Odds When X=1}}{\text{Odds When X=0}}.$$

Intuitively, it measures how the odds change when we move X from 0 to 1.

Step 3: Logit regression

Consider the following logit regression model:

$$\text{logit}(\text{Win}) = \log\left(\frac{\text{Win}}{1-\text{Win}}\right) = -1.40 + 0.33 \times \text{Hours}.$$

This model shows the relationship between **Hours** (the number of hours a party spends on campaign during the election) and **Win** (the probability of winning the election).

- When the party spends one additional hour on campaigns, we know
  - the corresponding change in log-odds of winning is 0.33.
  - the corresponding odds-ratio is $e^{0.33} \approx 1.39$.
- Should the party spend more time on campaigns?

$$e^{\beta} = \frac{\text{Odds When X=1}}{\text{Odds When X=0}}.$$

	It means	So more hours	Therefore
OR=1	(Odds when Hour=1) = (Odds when Hour=0)	do not change the odds (of winning)	Perhaps, but not sure if campaign helps (or not)
OR>1	(Odds when Hour=1) > (Odds when Hour=0)	increase the odds (of winning)	Campaign is a good idea
OR<1	(Odds when Hour=1) < (Odds when Hour=0)	reduce the odds (of winning)	Campaign is a bad idea

Given that $\text{OR}_{Hours} > 1$, the data support more hours on campaigns.

Live session teaser

- Draw on the 2011 Canadian National Election Study.
- Use packages `survey` or `srvyr` to create a survey object in `R` based on the survey design (e.g., strata and weights) – we have done this in Week 4.
- Use logit, such as `glm()` and `svyglm()` in `survey` package, to study the correlates of public attitudes against abortion.
- Use the `predict()` function in `R` to obtain predicted log-odds and probabilities.