Generalized Linear Model: Logit Regression

Advanced Topics in Quant Social Research

Plan for the day

- From OLS to GLM
  - Recap on the fundamentals of OLS
  - Why OLS can "fail" with binary/binomial outcomes
- Using logit regression to model binary/binomial outcomes
  - Treating outcomes as probability
  - Logit regression: Setup and inferences
- Concluding remarks
  - Wrapping up the loose ends
  - Other generalized linear models

From OLS to GLM: Recap on the fundamentals of OLS

- "Least squares linear regression" (aka **ordinary least squares**, OLS) uses a **linear** function to model the relationship between Xs (predictors) and Y (dependent variable).
  - A **bivariate** linear regression model (of one predictor) can be specified as
    $$Y = \alpha + \beta X + \epsilon,$$
    where α and β are the **intercept** (or constant) and the **slope** of the linear function.
  - A **multiple** linear regression (of k predictors) can be specified as
    $$\begin{aligned} Y &= \alpha + X\beta + \epsilon \\ &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \beta_k X_k + \epsilon, \end{aligned} \tag{1}$$
    where β_k is the corresponding slope of predictor X_k.
- We use ϵ to represent error or disturbance in the **data generation process** (DGP).

From OLS to GLM: Recap on the fundamentals of OLS

- With certain assumptions (linearity, independence, normality and homoscedasticity/equal variance of Y),
  - OLS can produce **unbiased** estimates for the intercept and slope(s)
  - The estimates are not necessarily **efficient**, as the variance/standard error of the estimates depends on statistical power (or sample size)
- OLS can also find the **"best"** fitted regression line based on our data by generating $\widehat{\alpha}$ and $\widehat{\beta}$s that lead to the smallest **sum of squared errors**
- OLS can model complicate non-linear DGP with the inclusion of **interaction** and **quadratic/polynomial** terms.

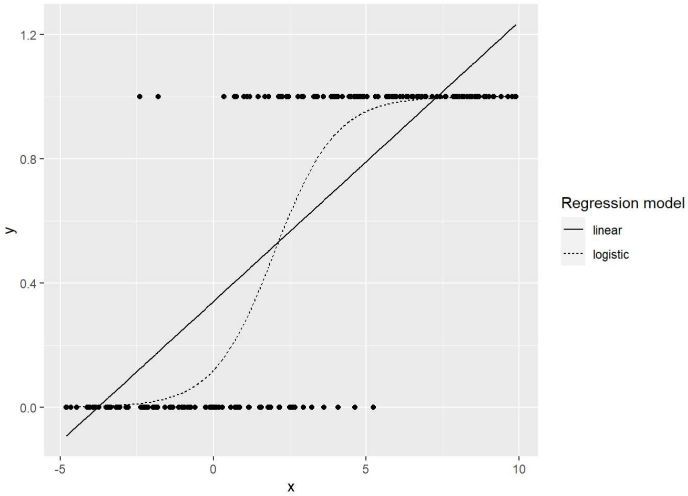
From OLS to GLM: When OLS goes wrong

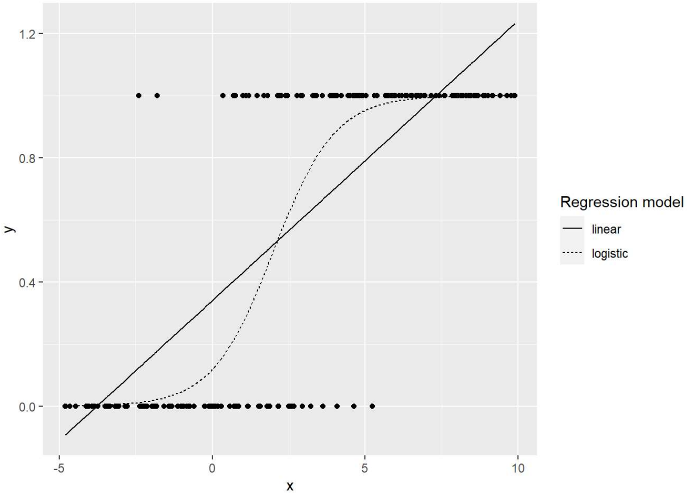
- OLS can go wrong when the dependent (or outcome or response) variable is a binary or binomial (Legler and Roback 2021).
  - **Binary** variables take on only two values: Success/yes ($Y = 1$) or fail/no ($Y = 0$).
  - **Binomial** variables the number of successes in n identical, independent trials with a constant probability p of success.
- Binary/binomial variables are ubiquitous in life. Using voter turnout as the example:
  - **Binary**: For individual voters, they can either vote or shirk (i.e., do not vote in the elections).
  - **Binomial**: For for all voters in a constituency, we can calculate the turnout rate (i.e., the percentage of eligible voters participating in the elections), if we assume each voter's decision to vote is independent and has the same probability.

From OLS to GLM: When OLS goes wrong

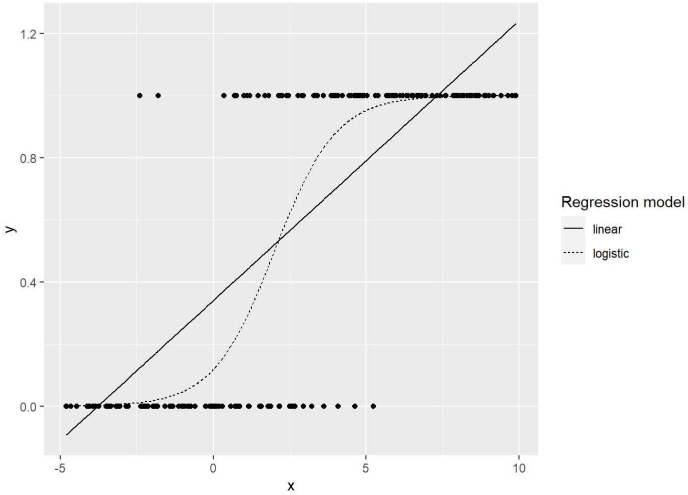
- **Unrealistic predicted outcome**. OLS can generate unrealistic predictions when Y is binary/binomial (or bounded).

- **Unequal variance of Y**. The homoscedasticity assumption will be be violated when Y is binary/binomial.


1.2
0.8
0.4
0.0
y
-5
0
5
10
x
Regression model
linear
logistic



Linear regression generate unrealistic predictions below 0 and above 1.



Linear regression generate unequal error across different values/levels of X. The error is larger as X gets close to the midpoint.

Using logit to model binary/binomial outcomes

- Using logit regression requires the following four steps:
  - Step 1: Treating binary/binomial outcomes as **probability**
  - Step 2: Transforming probability into **log odds**
  - Step 3: Using **linear function** to model log odds
  - Step 4: Interpreting the estimated slopes using **odds ratio**

Using logit regression: Treating outcomes as a probability

- Logit regression is used to model a **binary** outcome or a probability, which we denote as p.

- By the probability **axioms**,
  - A probability p can only take the values between 0 and 1.
  - The probability of an event happening (e.g., two countries fight) and that of the same event not happening (e.g., two countries do not fight) will sum up to 1.

- In practice, we use 0 and 1 for the outcome variable to indicate whether an event takes place

$$\begin{cases} Y = 1 & \text{when an event takes place} \\ Y = 0 & \text{when an event does not take place,} \end{cases} \tag{2}$$

  and we aim to use logit to estimate p, the probability of an event occuring (i.e., $P(Y = 1)$).

Logit regression: Transforming probability into log odds

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.

Logit regression: Transforming probability into log odds

If p is the probability of an event, then the odds is

$$\frac{p}{1-p},$$

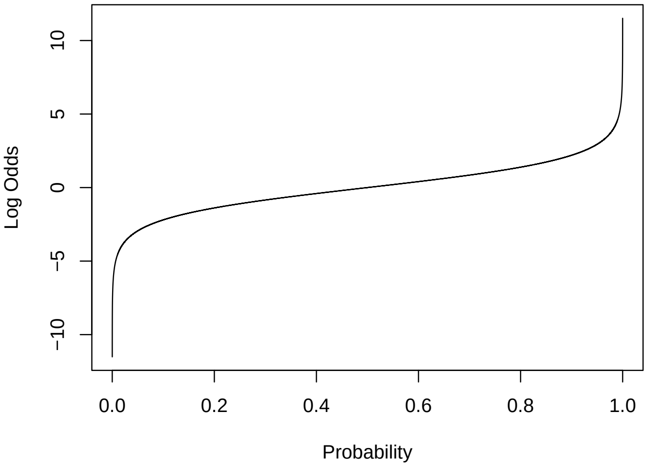
which suggests the chance of an event taking place relative to the opposite scenario.

- Say today's probability of raining is 0.8, such that $P(\text{rain}) = 0.8$.
- The **odds** of raining is then

$$\frac{P(\text{rain})}{P(\text{no rain})} = \frac{P(\text{rain})}{1 - P(\text{rain})} = \frac{0.8}{1-0.8} = \frac{0.8}{0.2} = 4.$$

- The **log of odds** is thus the natural log of 4

$$\log_e\left(\frac{P(\text{rain})}{P(\text{no rain})}\right) = \ln\left(\frac{P(\text{rain})}{P(\text{no rain})}\right) = \ln(4).$$



As we transform a binomial/binary variable from probability into log odds, it is no longer bounded between 0 and 1.

Logit regression: Using **linear function** to model log odds

- **Logit regression is a generalized linear model** (GLM) as we are using a **linear function** to model log odds.

- Logit regression first uses the **logit link function**:

$$\text{logit}(P(Y=1)) = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right),$$

where p is the probability that $Y = 1$.

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

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

Logit regression: Assumptions

- Logit regression is not assumption free, and some assumptions may not be plausible in real life.
  - The dependent variables can only **take the value of 0 or 1**.
  - The observations must be **independent** of each other.
  - The log of odds must be a **linear** function of the predictors.

Logit regression: Interpreting the estimated slopes

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

- **One-unit increase in** X (e.g., moving X from 0 to 1) corresponds to β changes in $\log\left(\frac{p}{1-p}\right)$

$$\begin{aligned} \beta &= \log(\text{Odds When X=1}) - \log(\text{Odds When X=0}) \\ &= \log\left(\frac{\text{Odds When X=1}}{\text{Odds When X=0}}\right). \end{aligned} \tag{3}$$

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

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

Example: Election campaigning and winning the elections

$$\text{logit}(P(\text{win})) = \log\left(\frac{P(\text{win})}{1 - P(\text{win}}\,)\right) = -1.40 + 0.33 \times \text{hours of TV campaign.}$$

- When the party spends one additional hour on campaigning on TV, we know
  - the corresponding **change in log-odds of winning** is 0.33.
  - the corresponding **odds-ratio** is

$$e^{0.33} \approx 1.39.$$

- Question: Should the party spend more time on televised 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 the odds ratio is larger than 1, the party should spend more hours on televised campaigns.

Concluding remarks: Wrapping up the loose ends

- Same as OLS, we can use the *p*-value and confidence intervals to test the statistical significance of the estimated odds ratio.
- Instead of using odds ratio to report your results, an alternative is to calculate predicted probabilities.
- Another alternative is to use **linear probability model**, but you will need to
  - Investigate the prevalence of implausible predicted outcome (i.e., $\widehat{Y} < 0$ or $\widehat{Y} > 1$)
  - Correct the standard errors using the **sandwich** estimator
- **[Extra]** Compared with OLS, logit and other non-linear GLMs are more likely to produce biased estimates when the model omits some important predictors and estimates across different models are not readily comparable (Breen et al 2018).

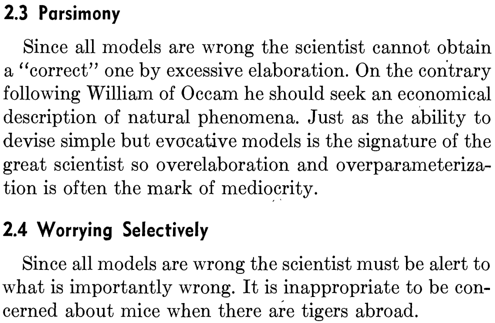
Concluding remarks: Other GLMs

- Varieties of modeling choices by the type of response variables.

Concluding remarks: Other GLMs

- Varieties of modeling choices by the type of response variables.
  - Continuous responses: OLS
  - Binary responses: Logit/probit
  - Ordinal responses: Ordered logit/probit
  - Categorical responses: 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.

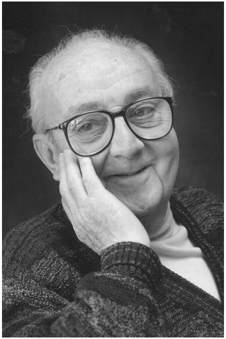


2.3 Parsimony

Since all models are wrong the scientist cannot obtain a “correct” one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

2.4 Worrying Selectively

Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.



Source: Box, George E.P. 1976. "Science and Statistics." *Journal of the American Statistical Association 71*(356): 791-799.

Key texts

- *Regression and Other Stories* (Vehtari et al 2020), Chapters 13-14
- *Beyond Multiple Linear Regression* (Legler and Roback 2021), Chapter 6
- "Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models," *Annual Reviews of Sociology* (Breen et al 2018)