MSBA7002 Business Statistics Tutorial 2

Yutao DENG

The University of Hong Kong

November 9, 2023

Table of Contents

Logistic Regression

Two type of statistical problems

- Regression Problem $\Leftarrow$ Y is continuous
  - Linear Regression

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p).$$

- Classification Problem $\Leftarrow$ Y is discrete/categorical
  - Probit Regression

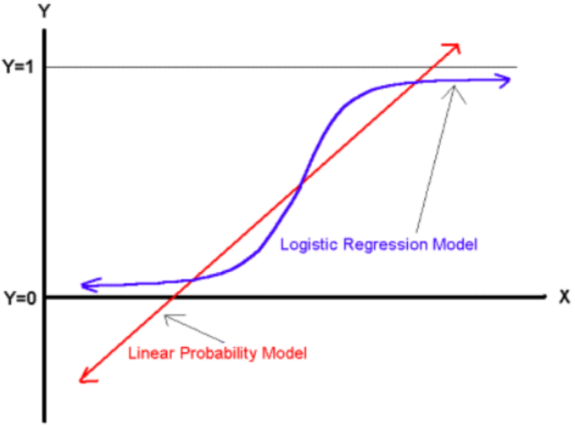
$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \begin{cases} \widehat{Y} = 1 & \mathrm{E}(\mathbf{Y}|\mathbf{X}; \widehat{\beta}) > 0.5 \\ \widehat{Y} = 0 & \mathrm{E}(\mathbf{Y}|\mathbf{X}; \widehat{\beta}) < 0.5 \end{cases}$$

  - Logistic Regression
  - Multiple level Logistic Regression (Generalization)

Logistic Regression

Why we need logistic regression?

- Probit Regression v.s. Logistic Regression



Logistic Regression

- Linear Regression

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p).$$

- "Generalized" Linear Model

$$f(\mathbf{Y}) = \mathbf{X}\beta + \varepsilon, \quad \mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p).$$

- log(Odds) $\implies f(\mathbf{Y})$

$$f(\mathbf{Y}) = \log\left(\frac{p(\mathbf{Y} = 1|\mathbf{X};\beta)}{1 - p(\mathbf{Y} = 1|\mathbf{X};\beta)}\right) = \log\left(\frac{p(\mathbf{Y} = 1|\mathbf{X};\beta)}{p(\mathbf{Y} = 0|\mathbf{X};\beta)}\right).$$

- $\implies p(\mathbf{Y} = 1|\mathbf{X};\beta) = \dfrac{e^{\mathbf{X}\beta}}{e^{\mathbf{X}\beta} + 1}.$

Logistic Regression

Three terminologies

- Odds (Odds ratio)

$$\text{Odds}(\mathbf{Y}|\mathbf{X};\beta) = \frac{p(\mathbf{Y}=1|\mathbf{X};\beta)}{1-p(\mathbf{Y}=1|\mathbf{X};\beta)} = \frac{p(\mathbf{Y}=1|\mathbf{X};\beta)}{p(\mathbf{Y}=0|\mathbf{X};\beta)}$$

- Logistic function

$$p(\mathbf{Y}=1|\mathbf{X};\beta) = \frac{e^{\mathbf{X}\beta}}{e^{\mathbf{X}\beta}+1}$$

- Likelihood

$$p(\mathbf{Y}=1|\mathbf{X};\beta)$$

Logistic Regression

Odds

- Odds (Odds ratio)
  - The conditional probability (likelihood) that the event will take place against the conditional probability (likelihood) that it will not.

$$\mathrm{Odds}(\mathbf{Y}|\mathbf{X};\beta) = \frac{p(\mathbf{Y}=1|\mathbf{X};\beta)}{1-p(\mathbf{Y}=1|\mathbf{X};\beta)} = \frac{p(\mathbf{Y}=1|\mathbf{X};\beta)}{p(\mathbf{Y}=0|\mathbf{X};\beta)}$$

- Formula of Odds ($\mathbf{x}_i = (1, x_{1i}, \ldots, x_{pi})$)

$$\pi_1(\mathbf{x}_i) = p(\mathbf{Y}_i = 1|\mathbf{X}_i = \mathbf{x}_i) = \frac{e^{\mathbf{x}_i\beta}}{e^{\mathbf{x}_i\beta}+1},$$

$$1-\pi_1(\mathbf{x}_i) = p(\mathbf{Y}_i = 0|\mathbf{X}_i = \mathbf{x}_i) = \frac{1}{e^{\mathbf{x}_i\beta}+1},$$

$$\mathrm{Odds}(\mathbf{Y}_i|\mathbf{x}_i;\beta) = \frac{\pi_1(\mathbf{x}_i)}{1-\pi_1(\mathbf{x}_i)} = e^{\mathbf{x}_i\beta}.$$

Logistic Regression

Logistic function

$$\log(\frac{p(\mathbf{Y}=1|\mathbf{X};\beta)}{p(\mathbf{Y}=0|\mathbf{X};\beta)}) = \mathbf{X}\beta + \varepsilon \iff p(\mathbf{Y}=1|\mathbf{X}) = \frac{e^{\mathbf{X}\beta}}{e^{\mathbf{X}\beta}+1}.$$

- Logistic function

$$f(\mathbf{X}) = \frac{L}{1+e^{-(\mathbf{X}-\mathbf{X}_0)\beta}} \quad \text{or} \quad f(x) = \frac{L}{1+e^{-k(x-x_0)}}$$

- Other S-shape functions
  - Sigmoid function $\iff$ Logistic function

$$f(x) = \frac{1}{1+e^{-x}}$$

  - Hyperbolic tangent function

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Logistic Regression
Likelihood

- Jointly Conditional Probability
- Likelihood of single observation

$$\pi_1(\mathbf{x}_i) = p(\mathbf{Y}_i = 1|\mathbf{X}_i = \mathbf{x}_i) = \frac{e^{\mathbf{x}_i\beta}}{e^{\mathbf{x}_i\beta}+1},$$
$$1-\pi_1(\mathbf{x}_i) = p(\mathbf{Y}_i = 0|\mathbf{X}_i = \mathbf{x}_i) = \frac{1}{e^{\mathbf{x}_i\beta}+1}.$$

- Likelihood (Joint Probability) of a particular event $\iff$ Joint Probability of repeated observations.

$$\mathcal{L}_i = \frac{n_i!}{y_i!(n_i-y_i)!}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}.$$

Logistic Regression
Likelihood

- Likelihood of this data (I unique events)

$$\mathcal{L} = \Pi_i^I \mathcal{L}_i = \Pi_i^I \frac{n_i!}{y_i!(n_i - y_i)!} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}.$$

- Log-likelihood

$$l = \ln \mathcal{L} = \sum_{i=1}^{I} \ln \mathcal{L}_i.$$

- Maximum Likelihood Estimation (MLE)

$$\frac{\partial l}{\partial \beta} = 0 \iff \frac{\partial l}{\partial \beta_0} = \frac{\partial l}{\partial \beta_1} = \ldots \frac{\partial l}{\partial \beta_p} = 0.$$

Logistic Regression

Deviance

```
Call:
glm(formula = HD ~ ., family = binomial, data = fram_data.f)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7051  -0.7268  -0.5556  -0.3329   2.4455

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.334797   1.036630  -9.005  < 2e-16 ***
AGE          0.062491   0.014995   4.167 3.08e-05 ***
SEXMALE      0.906102   0.157639   5.748 9.03e-09 ***
SBP          0.014838   0.003886   3.818 0.000135 ***
DBP          0.002875   0.007620   0.377 0.705941
CHOL         0.004459   0.001505   2.962 0.003053 **
FRW          0.005795   0.004055   1.429 0.152957
CIG          0.012309   0.006087   2.022 0.043150 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1469.3  on 1392  degrees of freedom
Residual deviance: 1343.1  on 1385  degrees of freedom
AIC: 1359.1

Number of Fisher Scoring iterations: 4
```

Logistic Regression

Deviance

- Deviance (of a model)

$$D_{\text{model}} = -2\ln\left(\frac{\mathcal{L}_{\text{model}}}{\mathcal{L}_{\text{perfect model}}}\right), \\ = -2\ln(\mathcal{L}_{\text{model}}) - 0, \\ = -2\ln(\mathcal{L}_{\text{model}}) = -2l_{\text{model}}.$$

- Null Deviance
  - deviance of null model, which only considers intercept.

$$D_{\text{null}} = -2l_{\text{null}}.$$

- Residual Deviance
  - deviance of fitted model

$$D_{\text{fitted}} = -2l_{\text{fitted}}.$$

Logistic Regression

Likelihood Ratio Test (LRT)

- H_0: There is no different between the deviance of fitted model fit2 and that of its reduced model fit1.
- H_1: There is significant difference (p $<$ 0.05)

$$D_{\text{fit2}} - D_{\text{fit1}} = -2\ln(\frac{\mathcal{L}_{\text{fit1}}}{\mathcal{L}_{\text{fit2}}}) \sim \chi^2(p),$$

- p is the difference of degree of freedom between fit1 and fit2.
- fit1: the shorter model; fit2: the longer model.
- anova(fit1,fit2,test = ”chisq”)

Logistic Regression
Summary

	Linear Regression	Logistic Regression
Method	OLS	MLE
Measure	Sum Square	Likelihood
Error of model	SSE	Deviance
Significance of β	t-test	z-test
Goodness of fit	F-test	χ^2-test

Table of Contents

Muliti-level Logistic Regression

Categorical Variables

- 2 levels
  - Logistic Regression
  - e.g. Weather prediction: Rainy v.s. Not Rainy
- more than 2 levels
  - Nominal Response
    - Nominal Logistic Regression
    - e.g. Weather prediction: Rainy v.s. Snowy v.s Sunny
  - Ordinal Response
    - Ordinal Logistic Regression
    - e.g. Grade Level Prediction: A v.s. B v.s C