Statistics for the Sciences

Logistic Regression

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Generalized Linear Models
- Logistic Regression

Generalized Linear Models

- Consider Y numerical response and covariates numerical or dummy variables.
- Recall linear models

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

  - $\mathbf{Y}$ is the vector of observed responses.
  - $\mathbf{X}$ the matrix of explanatory variables (also known as the design matrix).
  - β is the vector of coefficients (parameters) to be estimated.
    - $\mathbf{X}\beta$ is called the **systematic part**
  - ε is the vector of random errors, deviations of the observed responses from the expected responses given by the systematic part.

Generalized Linear Models

The linear model can be rearranged to the following tripartite form:

1. The random component: $\mathbf{Y}$ has independent Normal distribution with constant variance σ^2 and $E(\mathbf{Y}) = \boldsymbol{\mu}$.
2. The systematic component: covariates in the form of an $n \times (p+1)$ design matrix $\mathbf{X} = (\mathbf{1}, \mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_p}) = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{pmatrix}$ produce a linear predictor $\boldsymbol{\eta}$ given by
$$\boldsymbol{\eta} = \boldsymbol{X}\beta,$$
where β is a $(p+1) \times 1$ regression parameter vector.
3. The **link** between the random and systematic components is given by
$$\boldsymbol{\mu} = \boldsymbol{\eta}.$$

Generalized Linear Models

- Generalized linear models (GLM) generalize the classical linear models by allowing two extensions.
  - First, the distribution in part 1 comes from a family of distributions, called **exponential family**, which includes the normal distribution as a special case.
  - Secondly, the link between the random and systematic components is given by $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, where g is called the **link function** which is monotone and differentiable.

- **Random Component**: Probability distribution for **Y**
- **Systematic component**: Specifies explanatory variables in the form of a 'linear predictor':
$$\boldsymbol{\eta} = \boldsymbol{X}\beta$$
- **Link function**: Connects $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, where $E(\mathbf{Y}) = \boldsymbol{\mu}$.

Generalized Linear Models

To simplify the notations, we consider the relationships between the scalar random variables instead of using matrix notation.

- Ordinary regression: Normal
- Logistic regression: Bernoulli
- Poisson regression: Poisson

- Other possibilities: Binomial, Exponential, Gamma, Geometric ...

Generalized Linear Models

- (1) Systematic component is a regression-like equation

$$\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- (2) The link function g is monotone and differentiable.

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

  - The function $g(\mu)$ is strictly increasing in μ.
  - So μ is an increasing function of the Systematic component because the inverse of an increasing function is still increasing.

Generalized Linear Models

- For **linear models**,

$$\mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

- $E(Y) = \mu$
- The link is identity: $\eta = g(\mu) = \mu$
- $\mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

Generalized Linear Models

- For **Logistic Regression** (Binary Response)

$$g(\mu) = g\big(p(x)\big) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

where $0 < p(x) < 1$ is the population proportion.

  - $E(Y) = \mu = p(x)$
  - The logit link: $\eta = g(\mu) = \log \frac{\mu}{1-\mu}$
  - $\eta = \log \frac{\mu}{1-\mu} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

Generalized Linear Models

- For **Poisson Regression** (Count Response)

$$g(\mu) = g\big(\lambda(x)\big) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

where $\lambda(x)$ is the population average count.

  - $E(Y) = \mu = \lambda(x)$
  - The log link: $\eta = g(\mu) = \log(\mu)$
  - $\eta = \log(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

Generalized Linear Models

- GLMs are fitted to data by the method of **maximum likelihood (ML)**, providing not only estimates of the regression coefficients but also estimated asymptotic (i.e., large-sample) standard errors of the coefficients. The ML estimates can be found using an **IRLS** (Iteratively Re-Weighted Least Squares) algorithm.
- To test the null hypothesis

$$H_0 : \beta_i = 0, i = 0, 1, \ldots, k$$

we can compute the **Wald statistic**

$$Z_0 = \frac{\widehat{\beta}_i - 0}{SE(\widehat{\beta}_i)},$$

where $SE(\widehat{\beta}_i)$ is the asymptotic standard error of the estimated coefficient $\widehat{\beta}_i$. Under the null hypothesis, Z_0 follows a standard normal distribution.

Generalized Linear Models

- **Deviance** (McCullagh and Nelder, 1989) is measure of goodness-of-fit.
  - It compares the fitted model M_1 to a saturated model M_2 (larger value of likelihood) that perfectly fits the data.
  - Deviance Formula: $D =$ -2[log L(fitted model) - log L(saturated model)]
  - A lower deviance indicates a better fit.

- **Likelihood Ratio Test** which compares the fit of two nested models (reduced model versus full model) can be used to test a set of regression parameters. The test statistic is a deviance.
  - Test statistic = -2(log-likelihood reduced - log-likelihood full)

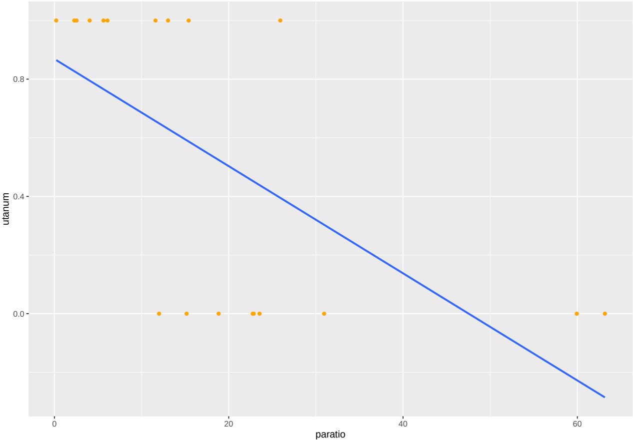
Logistic Regression

- Example (`polis.csv`): Polis et al. (1998) studied the factors that control spider populations on islands in the Gulf of California. We will use part of their data to model the presence/absence of lizards (Uta) against the ratio of perimeter to area (P/A, as a measure of input of marine detritus) for 19 islands in the Gulf of California.
  - response: `utanum`
  - predictor: `paratio`

```
##         island paratio uta utanum
## 1         Bota   15.41   P      1
## 2       Cabeza    5.63   P      1
## 3      Cerraja   25.92   P      1
## 4  Coronadito   15.17   A      0
## 5       Flecha   13.04   P      1
## 6     Gemelose   18.85   A      0
## 7     Gemelosw   30.95   A      0
## 8     Jorabado   22.87   A      0
## 9       Mitlan   12.01   A      0
## 10        Pata   11.60   P      1
## 11       Piojo    6.09   P      1
## 12       Smith    2.28   P      1
## 13     Ventana    4.05   P      1
## 14     Bahiaan   59.94   A      0
## 15     Bahiaas   63.16   A      0
## 16      Blanca   22.76   A      0
## 17    Pescador   23.54   A      0
```

Logistic Regression

- Can we use Linear Regression?



Logistic Regression

- Linear regression might produce probabilities less than zero or bigger than one. So it can not give a good estimate of $E(Y|X = x) = Pr(Y = 1|X = x)$. Logistic regression is more appropriate.
- Logistic regression uses the form

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- Furthermore,

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X.$$

  - This function of $p(X)$ is called the **logit** or **log odds** (by log we mean natural log : ln).

Logistic Regression with Several Predictors

- Suppose that there are p predictors: $X_1, \ldots, X_p$.
- Just like before

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

- And just like before

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Logistic Regression

- We use maximum likelihood to estimate the parameters

```
## # A tibble: 1 x 8
##   null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##           <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1          26.3      18  -7.11  18.2  20.1     14.2          17    19
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept)    3.61      1.70       2.13  0.0334    1.01     8.04
## 2 paratio       -0.220     0.101     -2.18  0.0289   -0.485   -0.0665
```

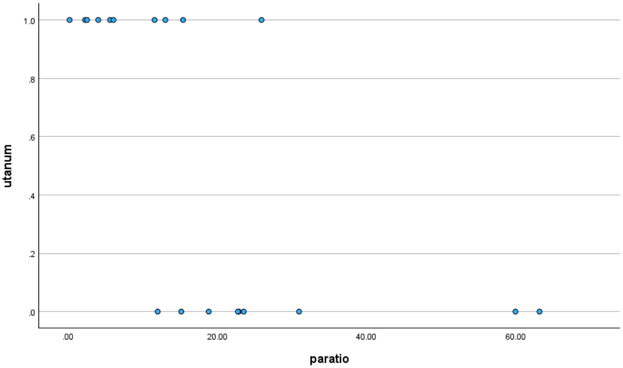
Logistic Regression

- Likelihood ratio test

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: utanum
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       18     26.287
## paratio  1   12.066        17     14.221 0.0005134 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

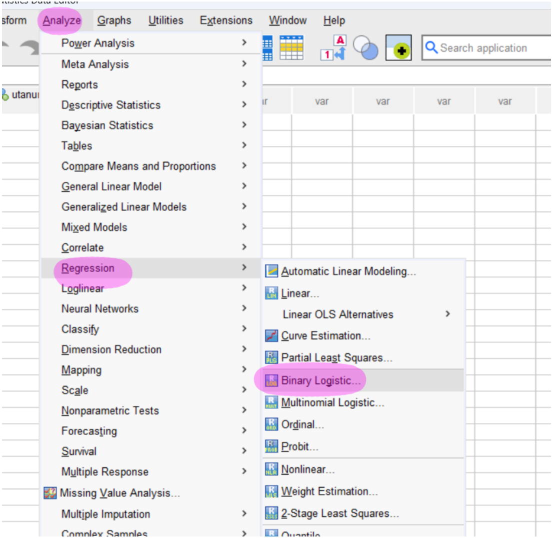
Lab

- Import the data `polis.csv`, you may get a scatter plot of the data

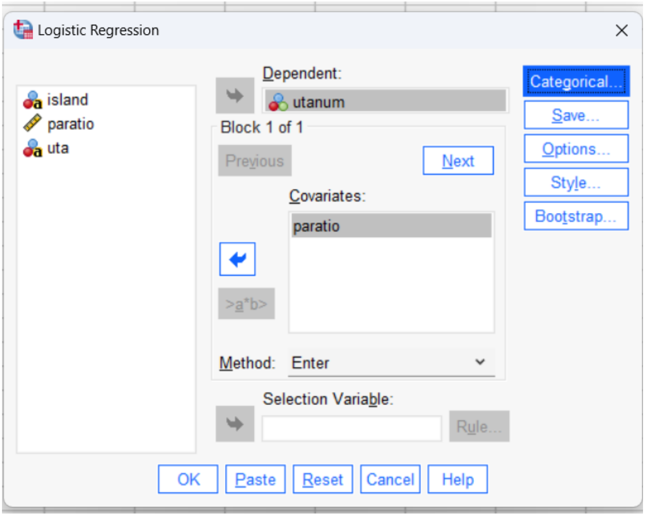


Lab

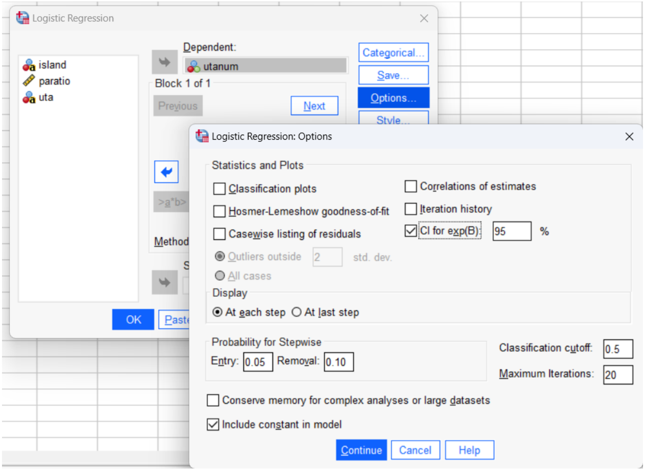
- Now let's fit a logistic regression



Lab


Logistic Regression
island
paratio
uta
Dependent:
utanum
Block 1 of 1
Previous
Next
Covariates:
paratio
>a*b>
Method:
Enter
Selection Variable:
Rule...
Categorical...
Save...
Options...
Style...
Bootstrap...
OK
Paste
Reset
Cancel
Help

Lab



Logistic Regression
island
paratio
uta
Dependent:
utanum
Block 1 of 1
Previous
Next
Method
OK
Paste
Categorical...
Save...
Options...
Style...
Logistic Regression: Options
Statistics and Plots
Classification plots
Hosmer-Lemeshow goodness-of-fit
Casewise listing of residuals
Outliers outside
2
std. dev.
All cases
Correlations of estimates
Iteration history
CI for exp(B):
95
%
Display
At each step
At last step
Probability for Stepwise
Entry:
0.05
Removal:
0.10
Classification cutoff:
0.5
Maximum Iterations:
20
Conserve memory for complex analyses or large datasets
Include constant in model
Continue
Cancel
Help

Lab

- By default, SPSS fits the probability of 1. See the `Dependent Variable Encoding` in the output

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Lab

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	14.221[a]	.470	.627

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Lab

- Since the log odds $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$ and $\exp(\beta_1) = e^{-0.22} = 0.803$. This means that For a one-unit increase in x, the odds of the outcome occurring are multiplied by 0.803.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	paratio	-.220	.101	4.771	1	.029	.803	.659	.978
	Constant	3.606	1.695	4.525	1	.033	36.821		

a. Variable(s) entered on step 1: paratio.

License