

Linear Regression (Review)

assumptions

1. linearity: $E[Y] = X\beta$ linear combination of the predictors

2. independence: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$
 $i \neq j$

3. Homoscedasticity: $\text{var}(Y_j) = \text{var}(\varepsilon_j) = \sigma^2$

4. Normality: $\varepsilon_j \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$

What if assumptions are not met? We can't use linear regression

case study: • response is not Normal
• constant variance (homoscedasticity) is not met

If the response is Binomial

$$Y_i \sim \text{Bin}(n, p_i) \quad i = 1, \dots, n$$

1. PMF $P(Y_i = y_i) = \binom{n}{y_i} p_i^{y_i} (1-p_i)^{n-y_i}$

2. mean $E[Y_i] = \mu_i = np_i$ from definition of expectation and variance for discrete

3. variance $\text{var}(Y_i) = \sigma_i^2 = np_i(1-p_i)$ v.v. 1

This variance can change for each y_i , so assumption 3 is not met.

On residuals vs fitted values we can see if assumptions are met. Here, not the case.

GLM (Generalized Linear Models)

if there is a correlation structure, mixed models can be used.

ACF of the AR(p)

$$\rho(h) - \phi_1 \rho(h-1) - \dots - \phi_p \rho(h-p) = 0$$

$\alpha_1, \dots, \alpha_r$: reciprocal roots

\downarrow
 m_1, \dots, m_r : multiplicity

$$\sum_{i=1}^r m_i = p$$

in this case

$$\rho(h) = \alpha_1^h p_1(h) + \dots + \alpha_r^h p_r(h)$$

$p_j(h)$ polynomial of order m

Exponential Family

Y is from exponential family if it can be written as :

$$f(y; \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

θ : the location (or natural) parameter

ϕ : the dispersion parameter

example : Binomial

$$Y \sim \text{Bin}(n, p)$$

$$\begin{aligned} f(y; n, p) &= P(Y=y; n, p) = \binom{n}{y} p^y (1-p)^{n-y} \\ &= \exp \left\{ \log \binom{n}{y} + y \log(p) + (n-y) \log(1-p) \right\} \\ &= \exp \left\{ \log \binom{n}{y} + y \log(p) + (n-y) \log(1-p) \right\} \\ &= \exp \left\{ \log \binom{n}{y} + \underbrace{y \log(p)}_{\theta} - \underbrace{y \log(1-p)}_{b(\theta)} + \underbrace{n \log(1-p)}_{c(y, \phi)} \right\} \\ &= \exp \left\{ \underbrace{y \log\left(\frac{p}{1-p}\right)}_{\theta} + \underbrace{n \log(1-p)}_{b(\theta)} + \underbrace{\log \binom{n}{y}}_{c(y, \phi)} \right\} \end{aligned}$$

Note that $\theta = \log\left(\frac{p}{1-p}\right)$, $\phi = 1$

$$\Leftrightarrow p = \frac{e^\theta}{1+e^\theta}$$

So the Binomial distribution is a member of the Exponential family.

$$\mu = E[Y] = b'(\theta)$$

$$\sigma^2 = \text{var}(Y) = b''(\theta) \phi$$

Introduction to Binomial Regression

Each y_i is a realization from a Y_i , Binomial

$$Y_i \sim \text{Bin}(1, p_i) \quad p_i \in [0, 1]$$

Binary Logistic Regression

$$Y_i \sim \text{Bin}(n_i, p_i)$$

assumption:

Y_i are independent

Goal: predict or explain p_i using covariates

$$x_{i,1}, \dots, x_{i,p}$$

Then predict $E[Y_i] = \mu_i = n_i p_i$

Systematic component:

$$\eta_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p}$$

how to link the linear predictor to the probability of success?

$$g(\mu_i) ?$$

Binomial Regression: link functions

1. Logit (logistic link):

$$\eta_i \stackrel{\text{set}}{=} g(p_i) = \log \left(\frac{p_i}{1-p_i} \right) = \theta \quad \text{natural parameter}$$

2. Probit:

$$g(p_i) = \Phi^{-1}(p_i)$$

Φ^{-1} : inverse $N(0,1)$ CDF

Binomial Regression Parameter Estimation

maximum likelihood estimation

n_i = number of trials

1. marginal p.m.f. :

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$$

$$p_i = \frac{e^{\theta_i}}{1 + e^{\theta_i}}, \quad \theta_i = \log\left(\frac{p_i}{1-p_i}\right) = \eta_i$$
$$= \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p}$$

2. joint p.m.f. :

because we have independence

$$f(y_i; n_i, p_i) = \prod_{i=1}^n \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$$
$$= \prod_{i=1}^n \binom{n_i}{y_i} \left(\frac{e^{\eta_i}}{1 + e^{\eta_i}} \right)^{y_i} \left(1 - \frac{e^{\eta_i}}{1 + e^{\eta_i}} \right)^{n_i - y_i}$$

3. likelihood function :

$$L(\beta) = \prod_{i=1}^n \binom{n_i}{y_i} \left(\frac{e^{\eta_i}}{1 + e^{\eta_i}} \right)^{y_i} \left(1 - \frac{e^{\eta_i}}{1 + e^{\eta_i}} \right)^{n_i - y_i}$$

4. log-likelihood function:

$$l(\beta) = \sum_{i=1}^n \left(y_i \eta_i - n_i \log(1 + e^{\eta_i}) + \log \binom{n_i}{y_i} \right)$$

not linear. We use iterative technique to maximize the log-likelihood function.

5. maximize!

Interpretation of Binomial Regression

$$\eta_i = \hat{\beta}_0 = \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} = \log \left(\frac{\hat{p}_i}{1 - \hat{p}_i} \right)$$

2 predictors: $x_{i,1}$ and $x_{i,2}$

we have ml estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$

Definition: Let event E have probability p of occurrence.
Then the odds in favor of E is:

$$O_E = \frac{p}{1-p}$$

biased coin: $P(H) = 3/4$ and $P(T) = 1/4$

$$O_E = \frac{p}{1-p} = \frac{(3/4)^2}{1 - (3/4)^2} = \frac{0.5625}{0.375} \approx 1.29$$

that we get H twice in a row.

It only holds for the logit link function.

β_0 : log odds of success when all predictors are equal to 0.

β_j : $X_{i,j}$ increase by 1 unit and all other predictors are held constant, the log-odds of success increases by β_j .

Odds of success increases by e^{β_j}

Binomial Regression in R

Occupancy : 0: not occupied
1: occupied