Linear Regression (Review)

assumptions

1. linearity: $E[Y] = X\beta$ — linear combination of the predictors
2. independence: $cov(\varepsilon_i, \varepsilon_j) = 0$ $\quad i \neq j$
3. Homoscedasticity: $var(Y_j) = var(\varepsilon_j) = 0$
4. Normality: $\varepsilon_j \overset{iid}{\sim} N(0, \sigma^2)$

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

$$Y_i \sim Bin(n, p_i) \qquad i = 1, \dots, n$$

1. PMF $P(Y_i = y_i) = \binom{n}{y} p_i^y (1-p_i)^{n-y}$
2. mean $E[Y_i] = \mu_i = np_i$
3. variance $var(Y_i) = \sigma_i^2 = np_i(1-p_i)$

from definition of expectation and variance for discrete r.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)

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

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

$\downarrow$

$m_1, \dots, m_r$: multiplicity $\quad \sum_{i=1}^{r} m_i = p$

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

$$Y \sim Bin(n,p)$$

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

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

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

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

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

$$\sigma^2 = var(Y) = b''(\theta)\,\phi$$

Introduction to Binomial Regression

Each y_i is a realization from a Y_i, Binomial

$$Y_i \sim Bin(1, p_i) \qquad p_i \in [0,1]$$

Binary Logistic Regression

$$Y_i \sim Bin(n_i, p_i)$$

assumption: Y_i are independent

<u>Goal</u>: 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:

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

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

$g(\mu_i)$?

Binomial Regression: link functions

1. Logit (logistic link):

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

2. Probit:

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

Φ^{-1}: inverse $N(0,1)$ CDF

Binomial Regression Parameter Estimation