Linear Regression (Review)

assumptions

1. linearity: $\mathbb{E}[Y] = X\beta$ 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: $\varepsilon_j \overset{iid}{\sim} N(0, \sigma^2)$

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

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

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

from definition of expectation and variance for discrete r.v.

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

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

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

θ : $y\log\left(\frac{p}{1-p}\right)$, $b(\theta)$: $n\log(1-p)$, $c(y,\phi)$: $\log\binom{n}{y}$

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.

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

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

Introduction to Binomial Regression

Each y_i is a realization from a Y_i, Binomial

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

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

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

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):

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

2. Probit:

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

Φ^{-1}: inverse $N(0,1)$ CDF

Binomial Regression Parameter Estimation

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

$$f(y_i; n_i, p_i) = \prod_{i=1}^{n} \binom{n_i}{p_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$$

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

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:

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

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

5. maximize!

Interpretation of Binomial Regression

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

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 occurence. 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 : $X_{:,j}$ 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

Temperature : in Celcius
Relative Humidity :
Light :
CO_2 measurement :

use "RCurl" package

"glm" function , family = "binomial"

format of the response : factor or
as a two-column matrix
(factor here)

Estimates computed using maximum likelihood estimation us IRWLS.

$\hat{\beta}_0 = -29.31$: . on the scale of the linear predictor (not exponentiated)

. average log-odds of an app. being occupied, everything else is 0, is about 29,3 %.

. $e^{\hat{\beta}_0} \approx 0$ (odd)

$\hat{\beta}_3$ (light) : . $e^{\hat{\beta}_3} \approx 1.02$ (odd)

. increase is multiplicative

$$e^{\hat{\eta}+1} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 (x_3+1) + \hat{\beta}_4 x_4}$$

$$= e^{\hat{\beta}_3} \underbrace{e^{\hat{\beta}_0 + \hat{\beta} x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4}}_{e^{\hat{\eta}}}$$

$$= e^{\hat{\beta}_3} e^{\hat{\eta}} \qquad e^{\hat{\beta}_3} \approx 1.02$$

Poisson Regression : model for count data

Y_i , $i = 1, \dots, n$ response variable, Poisson

$X_{i,j}$, $j = 1, \dots, p$ systematic component

$Y_i \overset{i}{\sim} Poi(\lambda_i)$

$$P(Y_i = y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

$y_i = 0, 1, 2, \dots$

$\lambda_i > 0$ potentially different for d.ff. measurement

mean : $\mu_i = E[Y_i] = \lambda_i$

variance : $\sigma_i^2 = var(Y_i) = \lambda_i$

canonical parameter : $\theta_i = \log(\mu_i) = \log(\lambda_i)$

canonical link function : $\log(\lambda)$

systematic component :

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

has to be positive. The mean (positive) should be linked to the linear predictor by a link function

$$g(\lambda_i) = \log(\lambda_i)$$ mean in a rate

$$\Rightarrow \quad \eta_i \overset{\text{set}}{=} g(\lambda_i) = \log(\lambda_i) = \theta \quad \left(\begin{matrix}\text{canonical}\\ \text{parameter}\end{matrix}\right)$$

$$\lambda_i = e^{\eta_i}$$

estimation by maximum likelihood.

Rate response : $\mu_i = \lambda_i = \frac{\text{count}}{\text{exposure time}} = \frac{y_i}{e_i}$

e_i : Known period of time

offset term : (log link)

$$g(\lambda_i) = \log(\lambda_i) = \log\left(\frac{y_i}{e_i}\right) = \log(y_i) - \log(e_i)$$

offset in glm formula. For different period of time of exposure.

family = "poisson"

Poisson Regression : parameter estimation

maximum likelihood estimation

1. marginal pmf :

$$P(Y_i = y_i) = \frac{e^{-\lambda_i} \lambda^{y_i}}{y_i!} \quad , y_i = 0, 1, 2, \ldots$$

2. joint pmf. :

$$f(y_i ; \lambda_i) = \prod_{i=1}^{n} \left(\frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \right)$$

3. likelihood function

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{e^{\lambda_i} \lambda_i^{y_i}}{y_i!} \right) \qquad \eta_i = \log(\lambda_i)$$

$$e^{\eta_i} = e^{\beta_0 + \beta_1 x_{i1} + \ldots}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{e^{-\eta_i}} e^{y_i \eta_i}}{y_i!} \right) = \prod_{i=1}^{n} \left(\frac{e^{y_i \eta_i - e^{\eta_i}}}{y_i!} \right)$$

4. log-likelihood function

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

5. maximize! using iterative techniques.

Interpreting the Poisson Regression model

remember : we use the log link $\log(\lambda_i) = \eta_i$

- β_0 : e^{β_0} can be interpreted as the mean of the response when each predictor is set to 0.
- β_j : e^{β_j} can be interpreted as the multiplicative increase in the mean of the response for a one unit increase in $x_{i,j}$, holding all other predictors constant.

$$y_i^{+1} = \hat{\lambda}_i^{+1} = e\{\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_j (x_{i,j} + 1) + \dots + \hat{\beta}_p x_{i,p}\}$$

$$= \exp\{\hat{\beta}_j\} \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_j (x_{i,j}) + \dots + \hat{\beta}_p x_{i,p}\}$$

$$= \exp\{\hat{\beta}_j\} \hat{\lambda}_i$$

Poisson Regression : Goodness of fit

Deviance : D

$$D = -2\underbrace{\ell(\hat{\beta})}_{\text{mle}}$$

$$\sum_{i=1}^{n}\left(y_i\eta_i - e^{\hat{\eta}_i} - \log(y_i!)\right)$$

a small deviance means a better fit.

The null deviance (deviance for the null model)

$$D_{null} = -2\sum_{i=1}^{n}\left(y_i\hat{\eta}_i - e^{\hat{\eta}_i} - \log(y_i!)\right)$$

$$\boxed{\text{if } \eta_i = \beta_0 \iff \hat{\lambda}_i = \bar{y}}$$

$$= -2\sum_{i=1}^{n}\left(y_i\log(\hat{\lambda}_i) - \hat{\lambda}_i - \log(y_i!)\right)$$

$$= -2\sum_{i=1}^{n}\left(y_i\log(\bar{y}) - \bar{y} - \log(y_i!)\right)$$

The saturated deviance (deviance for the saturated model)

A parameter for each data point.

$$D_{sat} = -2 \sum_{i=1}^{n} \left(y_i \log(y_i) - y_i - \log(y_i!) \right)$$

The residual deviance

D_p is the model that we are interested

$$\begin{aligned} D_{resid} &= D_p - D_{sat} \\ &= -2 \sum_{i=1}^{n} \left(y_i \log(\hat{\lambda}_i) - \hat{\lambda}_i - \log(y_i!) \right) \\ &\quad + 2 \sum_{i=1}^{n} \left(y_i \log(y_i) - y_i - \log(y_i!) \right) \\ &= \dots \\ &= 2 \sum_{i=1}^{n} \left(y_i \log\left(y_i / \hat{\lambda}_i \right) - (y_i - \hat{\lambda}_i) \right) \end{aligned}$$

This quantity has a Chi-2 distribution $(n-(p+q))$ degrees of freedom

$$\chi^2_{n-(p+1)}$$

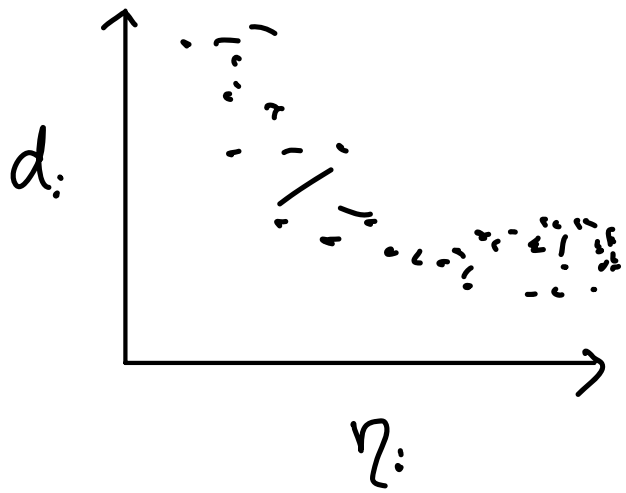
Goodness of fit test

H_0 : The model with p parameters fits the data well

Deviance residuals

$$d_i = \text{sign}(y_i - \hat{\lambda}_i)\sqrt{2\left(y_i\left(\log(y_i/\hat{\lambda}_i) - (y_i - \hat{\lambda}_i)\right)\right.}$$

$d_i \sim N(0,1)$. under H_0 and with a high number of counts.



. if we see a nonlinear trend, this could indicate lack of fit

Other Goodness of Fit

Pearson's χ^2

$$\chi^2 = \sum_{i=1}^{n} (O_i - E_i)^2 / E_i$$

$$= \sum_{i=1}^{n} \underbrace{(y_i - \hat{\lambda}_i)^2 / \hat{\lambda}_i}_{\text{Squared standardized residual}}$$

$$\sim \chi^2_{(n-p+1)}$$

A large χ^2 is evidence against the null that the model fit is sufficient.

$$p_i = \frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}} \qquad \text{pearson residual}$$

Overdispersion

$E[Y_i] = \lambda_i \qquad \text{var}(Y_i) = \lambda_i \quad \text{if } Y_i \text{ is Poisson}$

if $\text{var}(Y_i) > E[Y_i]$, we have overdispersion

Real overdispersion (vs) apparant overdispersion

- zero inflation
  ⇒ mixture models or hurdle models
  e.g. some insurance claim models.
- spatial or temporal correlation

- outliers that are "true measurements"
  ⇒ detecting outliers and removing them
- missing predictors
- inappropriate link function

How to detect ?

$$\mathbb{E}[Y_i] = \lambda_i \quad , \quad \text{var}(Y_i) = \lambda_i \phi$$

ϕ is the dispersion parameter.

$$\hat{\phi} = \frac{\sum_{i=1}^{n} (y_i - \hat{\lambda}_i)^2 / \hat{\lambda}_i}{n - (p+1)} \sim \chi^2_{n-(p+1)}$$

p : number of predictors in the model

if statistically significant, then we have evidence of overdispersion.

Solutions (to overdispersion)

1. Quasi-likelihood methods

$$\hat{se}(\hat{\beta}_j) = \hat{se}(\hat{\beta}_j)\sqrt{\hat{\phi}}$$

- only the estimated variances will be affected.

2. Negative Binomial regression