Week 2: Generalized Linear Models

MATH-516 Applied Statistics

Linda Mhalla

2024-02-28

Section 1

Introduction

Introduction

- Linear models are only suitable for data that are (approximately) normally distributed
- However, there are many settings where we may wish to analyse a response variable which is not necessarily continuous, including when
- Y is **binary**
- Y is a **count** variable
- Y is **continuous, but non-negative**
- We consider particular distributions for binary/proportion and counts data, in order to do likelihood-based inference

Exponential Family

Definition. The distribution of Y is of exponential type if its density can be written as

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

where $\theta \in \mathbb{R}$ is the canonical parameter, $\varphi \in (0, \infty)$ is the dispersion parameter, and b, c are real functions.

If $b \in C^2$, it can be shown using the moment generating function $m(t) = \mathbb{E}e^{tX}$ that

- $\mu := \mathbb{E}(Y) = b'(\theta)$
- $\mathrm{var}(Y) = \varphi b''(\theta)$
- $\mathrm{var}(Y) = \varphi V(\mu)$, where V is called variance function

Gaussian Distribution

$$\begin{aligned} f(x, \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for} \quad x, \mu \in \mathbb{R} \text{ and } \sigma^2 \in (0, \infty) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{x^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{x\mu - \mu^2/2}{\sigma^2} + \left[-\frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right]\right) \end{aligned}$$

hence

- $b(\theta) = \mu^2/2$ and $c(x, \sigma^2) = -\frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$ with $\theta = \mu$ and $\varphi = \sigma^2$
- $\mathrm{var}(Y) = \varphi \cdot 1 \Rightarrow V(\mu) \equiv 1$ (variance does not depend on expectation)

Bernoulli Distribution

$$
\begin{aligned}
f(x,p) &= p^x(1-p)^{1-x} \quad \text{for} \quad x \in \{0,1\} \text{ and } p \in (0,1) \\
&= \exp\left\{x \log p + (1-x)\log(1-p)\right\} \\
&= \exp\left\{x \log \frac{p}{1-p} + \log(1-p)\right\}
\end{aligned}
$$

hence

- $\theta = \log \frac{p}{1-p}$, $\varphi = 1$, $b(\theta) = -\log(1-p)$, and $c(x,\varphi) = 0$
- $\mathrm{var}(Y) = p(1-p)$ and $\mu = \mathbb{E}X = p \Rightarrow V(\mu) = \mu(1-\mu)$

Poisson Distribution

$$\begin{aligned} f(x) &= \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{for} \quad x \in \{0, 1, 2 \ldots\} \text{ and } \lambda \in (0, \infty) \\ &= \exp\left(x \log \lambda - \lambda + \log(1/x!)\right) \end{aligned}$$

hence

- $\theta = \log \lambda$, $\varphi = 1$, $b(\theta) = e^{\theta}$, and $c(x, \varphi) = \log(1/x!)$
- $\mathrm{var}(Y) = \lambda$ and $\mu = \mathbb{E}X = \lambda \Rightarrow V(\mu) = \mu$

Section 2

GLMs

Generalized Linear Models

- Generalized linear models (GLMs) combine a model for the conditional mean with a distribution (usually within the exponential family) for the response variable and a link function tying predictors and parameters
  - Linear regression (with normal errors) is a special case of a generalized linear model
- Today, we will give an introduction to generalized linear models and focus in particular on Poisson regression
  - We will only discuss the case of independent observations
  - Extensions of generalized linear models for correlated and longitudinal (the so-called **generalized linear mixed models**), will be covered in few weeks

Notations

- The starting point is the same as for linear regression:
  - We have a random sample of independent observations

$$(Y_i, \mathrm{X}_{i1}, \dots, \mathrm{X}_{ip}), \quad i = 1, \dots, N$$

  where Y is the response variable and $\mathrm{X}_1, \dots, \mathrm{X}_p$ are p explanatory variables or covariates which are assumed fixed (non-random)

- The goal is to model the response variable as a function of the explanatory variables
- Let μ_i denote the (conditional) mean of Y_i given covariates,

$$\mu_i = \mathbb{E}(Y_i \mid \mathrm{X}_{i1}, \dots, \mathrm{X}_{ip})$$

- Let η_i denote the linear combination of the covariates that will be used to model the response variable,

$$\eta_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip}$$

Definition

- There are three building blocks to the generalized linear model:
    - A probability distribution for the outcome Y that is a member of the exponential family (normal, binomial, Poisson, gamma, inverse gaussian, …)
    - A linear predictor $\eta = \mathbf{X}^\top \beta$
    - A function g, called link function, that links the mean of Y_i to the predictor variables, $g(\mu_i) = \eta_i$
- The link between the mean of Y and the regression "line' ' is

$$g\left\{\mathbb{E}(Y \mid \mathrm{X}_1, \ldots, \mathrm{X}_p)\right\} = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_p \mathrm{X}_p$$

Link Function

- The link function connects the mean to the explanatory variables

$$\begin{aligned} g(\mu_i) = \eta_i &= \beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip} \\ \Leftrightarrow \quad \mu_i = g^{-1}(\eta_i) &= g^{-1}(\beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip}). \end{aligned}$$

- In the ordinary linear regression model, we do not impose constraints on the mean μ_i and $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{X}_{i1} + \cdots + \hat{\beta}_p \mathrm{X}_{ip}$ can take on any value in $(-\infty, \infty)$
- For some response variables, we would need to impose constraints on the mean
  - For Bernoulli responses, the mean $\mu = p$ must lie in the interval $(0, 1)$
  - For Poisson responses, the mean λ must be positive
- An appropriate choice of link function sets μ_i equal to a transformation of the linear combination η_i so as to avoid any parameter constraints on β

Choice of Link Function

Certain choices of the link function facilitate interpretation or make the likelihood function convenient for optimization

- For the Bernoulli and binomial distributions, an appropriate link function is the logit function

$$\mathrm{logit}(\mu) := \log\left(\frac{\mu}{1-\mu}\right) = \eta \quad \Leftrightarrow \quad \mu = \frac{\exp(\eta)}{1+\exp(\eta)}$$

- For the Poisson distribution, an appropriate link function is the natural logarithm

$$\log(\mu) = \eta \quad \Leftrightarrow \quad \mu = \exp(\eta)$$

- For the normal distribution, an appropriate link function is the identity function, $\mu = \eta$

MLE in GLM

- $\ell(\beta) = \sum_n \frac{Y_n \theta_n - b(\theta_n)}{\varphi} + c(\varphi, Y_n)$, where
  - $\theta_n = (b')^{-1}(\mu_n)$ and $\mu_n = g^{-1}(\mathbf{X}_n^\top \beta)$

$\Rightarrow$ maximization done via Iteratively Reweighted Least Squares (IRLS) (requires gradient vector and Hessian matrix)

- $U_n(\beta) := \frac{1}{\varphi} w_n g'(\mu_n)(Y_n - \mu_n) X_n$, with $w_n = [V(\mu_n)\{g'(\mu_n)\}^2]^{-1}$
  - shown using the chain and inverse function rules
- Fisher information: $\mathbf{I} = \frac{1}{\varphi}\mathbb{E}(\mathbf{X}^\top \mathbf{W} \mathbf{X})$
  - weight matrix $\mathbf{W}$ diagonal with weights w_n
  - log-likelihood is concave and IRLS converges to the MLE
  - one can work with the Hessian (full Newton) instead of the expected Hessian (Fisher scoring): beware of negative weights!

See Section 3.1 in Wood's book

MLE in GLM

MLE asymptotic theory implies that

- $\hat{\beta} \to \mathcal{N}_p(\beta, \mathbf{I}^{-1})$ [Wald]
    - φ is hidden. If unknown, estimate it consistently and use Cramer-Slutzsky
    - tests for subsets of β are based on the corresponding marginal normal distributions (provided by `summary(glm)` in `R`)
    - can be used to obtain CIs. Use `confint.default(glm, level=.95)` in `R`
- Let $H_0 : \beta_{p-m+1} = ... = \beta_p = 0$ hold in the GLM, $\hat{\beta}$ denotes parameter estimates in the model, and $\tilde{\beta}$ denotes parameter estimates in the submodel given by the linear constraints in H_0. Then [likelihood ratio]

$$2\{\ell(\hat{\beta}) - \ell(\tilde{\beta})\} \to \chi^2_m$$

    - can only be used when φ is known. Use `car::Anova(glm)` in `R`
    - can be used to get CIs (inverting the acceptance region) and are preferred to Wald's CIs. Use `confint(glm, level=.95)` in `R`

Deviance

Definition

1. The saturated model is a model with the largest possible amount of parameters (i.e., $p = N$ and $\mu_n = y_n$)
2. The statistic $D(\mathbf{Y}, \widehat{\beta}) = 2\varphi\{\widehat{\ell}(\mathbf{Y}) - \ell(\widehat{\beta})\}$, where $\widehat{\ell}(\mathbf{Y})$ denotes the maximized log-likelihood of the saturated model, is called the deviance

- it is a goodness-of-fit measure
  - for linear model, it is equal to the residual sum of squares R^2
- it measures the discrepancy in fit between the full and the fitted model and $\varphi^{-1} D(\mathbf{Y}, \widehat{\beta}) \overset{\cdot}{\sim} \chi^2_{N-p-1}$ if the fitted model is adequate ($p+1$ is the number of β's, including the intercept)
- model `summary(glm)` in `R` provides:
  - null deviance: deviance of the intercept-only model ($N-1$ df)
  - residual deviance: deviance of the provided model ($N-p-1$ df)
- can be used for model comparison when φ is unknown (F statistic)

Model Checking: Residuals

- Pearson residuals, a.k.a. standardized residuals

$$\epsilon_n^p = \frac{y_n - \hat{\mu}_n}{\sqrt{V(\hat{\mu}_n)}}$$

$\Rightarrow$ no trend in mean nor variance when plotted against fitted values

  - departure is proof against linearity
  - are obtained by `residuals(glm, type="pearson")`
  - should have zero mean but distribution can be asymmetric around 0

- Deviance residuals

$$\epsilon_n^d = sign(y_n - \hat{\mu}_n)\sqrt{d_n},$$

where $D(\mathbf{Y}, \hat{\beta}) = \sum_{n=1}^{N} d_n \Rightarrow$ expected to behave like $\mathcal{N}(0, \varphi)$ (if the model holds)

  - departure is proof against response distribution
  - are obtained by `residuals(glm) = residuals(glm, type="deviance")`

Section 3

Poisson Regression

Poisson Regression

- Poisson regression assumes that the outcome variable Y_n follows a Poisson distribution with parameter μ_n, $Y_n \sim \text{Po}(\mu_n)$, where $\mu_n = \mathbb{E}(Y_n) = \text{var}(Y_n)$
- We use $\log(x)$ as link function to ensure positivity of the mean,

$$g\{\mathbb{E}(Y_n)\} = g(\mu_n) = \log\{\mathbb{E}(Y_n)\} = \beta_0 + \beta_1 \mathrm{X}_{n1} + \cdots + \beta_p \mathrm{X}_{np}$$

- Equivalently, we could say that the outcome for individual n, Y_n, follows a Poisson distribution with mean

$$\mathbb{E}(Y_n) = \mu_i = \exp(\beta_0 + \beta_1 \mathrm{X}_{n1} + \cdots + \beta_p \mathrm{X}_{np})$$

Coefficient interpretation for β_k in Poisson regression

- Let $\mathbf{x}$, $\mathbf{x}_+$ be two vectors which differ only in their kth components, respectively x_k and $x_k + 1$
- When $\mathbf{X} = \mathbf{x}$, the model linking the mean to the variable Y is

$$\mu_n(\mathbf{x}) = \mathbb{E}(Y_n \mid \mathbf{X} = \mathbf{x}) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)$$

whereas, when $\mathbf{X} = \mathbf{x}_+$, we have

$$\mu_n(\mathbf{x}_+) = \mathbb{E}(Y_n \mid \mathbf{X} = \mathbf{x}_+) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j + \beta_k\right)$$

- The ratio between the two means, $\mu_n(\mathbf{x}_+)/\mu_n(\mathbf{x})$, is $\exp(\beta_k)$
- When X_k increases by one unit, the mean of Y is **multiplied** by $\exp(\beta_k)$

Example

Daily air quality measurements in New York, May to September 1973

```r
glm_poisson <- glm(Ozone ~ Solar.R + Temp + Wind, data = ozone,
                  family = "poisson", subset = trainset)
# car::Anova(glm_poisson, type="3") : for LRTs
summary(glm_poisson)
```

```
Call:
glm(formula = Ozone ~ Solar.R + Temp + Wind, family = "poisson",
    data = ozone, subset = trainset)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.7551989  0.2253574   3.351 0.000805 ***
Solar.R      0.0021881  0.0002298   9.520  < 2e-16 ***
Temp         0.0407351  0.0023818  17.103  < 2e-16 ***
Wind        -0.0822691  0.0058413 -14.084  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2090.98  on 88  degrees of freedom
Residual deviance:  581.99  on 85  degrees of freedom
AIC: 1058.2

Number of Fisher Scoring iterations: 4
```

Example

- The residual deviance should be close to the degrees of freedom (85), which is not the case here!
    - This could be a result of overdispersion where the variation is greater than predicted by the model
- Checking overdispersion

```r
mean(ozone$Ozone)
```

```
[1] 42.0991
```

```r
var(ozone$Ozone)
```

```
[1] 1107.29
```

```r
var(ozone$Ozone)/mean(ozone$Ozone)
```

```
[1] 26.30199
```

Negative Binomial

- Another alternative to the classical Poisson regression when data are over-dispersed is the negative-binomial regression
- In the negative-binomial distribution, the mean is identical to that of the Poisson while the variance is

$$\mathrm{Var}(Y) = \mu + \frac{\mu^2}{\theta}$$

- As θ increases, the variance approaches the mean (more like the classical Poisson distribution)
- The overdispersion in the negative binomial makes it a good candidate for modelling gene expressions (highly variable)

Negative Binomial

```r
library(MASS)
# Negative binomial regression
glm_nb <- glm.nb(Ozone ~ Solar.R + Temp + Wind, data = ozone, subset = trainset)
summary(glm_nb)
```

```

Call:
glm.nb(formula = Ozone ~ Solar.R + Temp + Wind, data = ozone,
    subset = trainset, init.theta = 6.346750802, link = log)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.6616379  0.5515469   1.200 0.230293
Solar.R      0.0019795  0.0005324   3.718 0.000201 ***
Temp         0.0409155  0.0059763   6.846 7.58e-12 ***
Wind        -0.0690672  0.0150123  -4.601 4.21e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(6.3468) family taken to be 1)

    Null deviance: 292.100  on 88  degrees of freedom
Residual deviance:  93.382  on 85  degrees of freedom
AIC: 734.06
```

Negative Binomial vs Poisson

LR test statistic has a non-standard distribution, even asymptotically

```
# P-value of the likelihood ratio test H0: theta=inf vs H1: th
lrtstat <- 2*as.numeric(logLik(glm_nb)-logLik(glm_poisson))
pchisq(lrtstat, df = 1, lower.tail = FALSE)/2
```

```
[1] 3.415969e-73
```

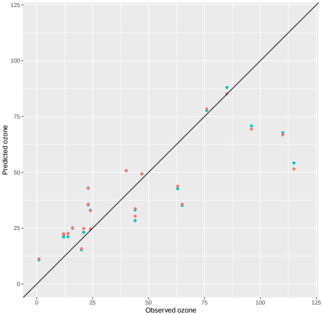
```
#ratio of deviance to dof
deviance(glm_nb)/ df.residual(glm_nb)
```

```
[1] 1.098611
```

Prediction

```r
# prediction on respone scale
pred.poisson <- predict(glm_poisson, newdata = ozone[testset, ], type = "response")
pred.nb      <- predict(glm_nb, newdata = ozone[testset, ], type = "response")

# store model predictions in a data frame
ozone.test <- data.frame("pred_poisson"= pred.poisson, "pred_nb"= pred.nb,
                         "Obs"=ozone[testset,"Ozone"])
library(ggplot2)
ggplot(ozone.test,aes(x=Obs,y=pred_poisson, col="red")) + geom_point() + geom_point
```



Offsets and comparison of counts

- Up to now, we implicitly assumed that the count variables Y were **comparable** between observations
  - the number of work accidents seen in a business in a given time period depends on the number of its employees
  - the number of cancer incidence per region depends on the number of inhabitants

If the counts are not comparable, we can compare the **rates** instead

- the work accident rate (number of accidents per employee)
- the chocolate chip rate (number of chocolate chips per square centimetre of a cookie)

Example: Car Accident

The National Highway Traffic Safety Administration (NHTSA) compiles statistics about road traffic deaths in the Fatality Analysis Reporting System. The yearly mortality counts for 2010 and 2018 are given in `crash` according to whether the accident occurred during daytime or nightime (`time`), and according to the NHTSA-defined geographic area (`region`)

- Let Y_n denotes the number of death in a given year in region n
- Let N_n denotes the number of inhabitants in region n

The goal is to estimate the relationship between the number of fatal car crash and timing of the incident

Example: Car Accident

- If we ignore the size of the population, the Poisson regression model (or negative binomial) would be

$$\log(\mu_n) = \log\{\mathbb{E}(Y_n)\} = \beta_0 + \beta_1\texttt{time} + \beta_2\texttt{year}$$

- If we want to account for the size of the population in a given region, we would model Y_n/N_n instead of Y_n. This amounts to setting

$$\log\left\{\frac{\mathbb{E}(Y_n)}{N_n}\right\} = \beta_0 + \beta_1\texttt{time} + \beta_2\texttt{year}$$

or equivalently

$$\log\left\{\mathbb{E}Y_n\right\} = \beta_0 + \beta_1\texttt{time} + \beta_2\texttt{year} + \log(N_n)$$

- The term $\log(N_n)$ is called an **offset** since it is included as a covariate, but there is no β coefficient to estimate (unity)

Parameter Interpretation with Offset

```
Call:
MASS::glm.nb(formula = ndeath ~ time + year + offset(log(popn)),
    data = crash, init.theta = 15.43950913, link = log)

Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -10.90616    0.07089 -153.854  < 2e-16 ***
timenight     0.22662    0.08164    2.776  0.00551 **
year2018      0.22997    0.08164    2.817  0.00485 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(15.4395) family taken to be 1)

    Null deviance: 56.064  on 39  degrees of freedom
Residual deviance: 40.269  on 37  degrees of freedom
AIC: 527.59

Number of Fisher Scoring iterations: 1

              Theta:  15.44
          Std. Err.:  3.51

 2 x log-likelihood:  -519.585
```

Parameter Interpretation with Offset

- The deviance statistic is 40.269 for 37 degrees of freedom (ratio of 1.0884). The corresponding p-value is 0.327, so there is no evidence that our fitted model is inadequate!
- In this setting, $\exp(\hat{\beta}_0) = \exp(-10.9062)$ corresponds to the estimated mortality rate during daytime in 2010, which is $1.83/100000$, i.e., a rate of 1.83 per $100\,000$ inhabitants (with 95% confidence interval $[1.60 \times 10^{-5}, 2.12 \times 10^{-5}]$)
- There is a $\exp(0.23)$ change in mortality from 2010 to 2018, corresponding to a 26% increase in road casualties