UC Irvine ISI-BUDS Day 12

Zhaoxia Yu

7/26/2022

Study Goals

Study Goals

- Review of LM
- GLM
  - Logistic Regression
  - Poisson Regression
  - Multinomial Regression
- The assumption of independent observations

Review of LM

A Linear Model (LM)

- Suppose

$$Y = \beta_0 + x_1 \times \beta_1 + \ldots + x_p \times \beta_p + \epsilon,$$

  where
  - the regressand Y is the response / outcome / dependent / endogenous variable
  - the regressors $(x_1, \cdots, x_p)$ are the p covariates / independent / explanatory variables
  - the random term ϵ has a zero mean and variance $\sigma^2 > 0$
  - the intercept is β_0, the other p coefficients are $\beta_1, \cdots, \beta_p$

A Linear Model (LM)

- Consider the ith observation:

$$Y_i = \beta_0 + x_{i1} \times \beta_1 + \ldots + x_{ip} \times \beta_p + \epsilon_i, i = 1, \ldots, n$$

- Basic assumptions
  - $E(\epsilon_i) = 0$, which is equivalent to $E(Y_i|X_i) = \beta_0 + x_{i1} \times \beta_1 + \ldots + x_{ip} \times \beta_p$
  - $Var(\epsilon_i) = \sigma^2$. Note, this is equivalent to say $Var(Y_i|X_i) = \sigma^2$.
  - $(\epsilon_1, \cdots, \epsilon_n)$ are mutually independent
- If $(\epsilon_1, \cdots, \epsilon_n)$ are i.i.d. $N(0, \sigma^2)$, we can derive t-tests and F-tests
- Question: what if the assumptions are violated?

Logistic Regression

A Motivating Example of GLM

- A motivating example: Consider a binary response variable, i.e., Y_i takes values of 0 or 1.
- Is LM a good choice for this problem?

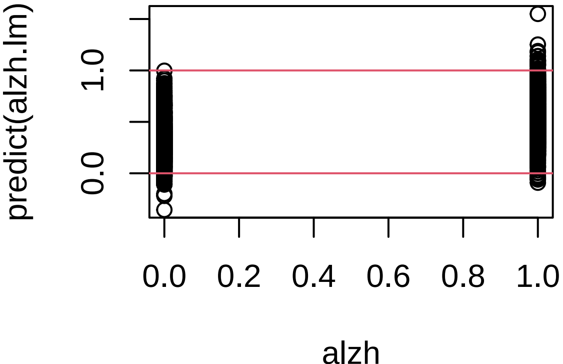
A Motivating Example of GLM (continued)

- Consider the Alzheimer data
- We create a binary variable

```
alzheimer=read.csv("alzheimer_data.csv", header = TRUE)
#dim(alzheimer)
#names(alzheimer)
attach(alzheimer)
#length(unique(id))
alzh=(diagnosis>0)*1 #"*1" to create a 0-1 variable
```

A Motivating Example of GLM (continued)


```
alzh.lm = lm(alzh ~ age + female + educ+lhippo + rhippo)
par(mar = c(4, 4, 0.5, 0.5))
plot(alzh, predict(alzh.lm)); abline(h=c(0,1), col=2)
```



A Motivating Example of GLM (continued)

- Is alzh.lm a good model for alzh?

- Several assumptions of the LM have been violated, and
- The predicted values using LM are not between 0 and 1!
- Let $X_i = (x_{i1}, \cdots, x_{ip})^T$, i.e., the vector of covariates for the ith subject.
- Let $\pi_i = E(Y_i|X_i)$, the expected probability. We would like to make sure that $\pi_i \in [0, 1]$
- How?

Logistic Regression

- Consider the a special transformation of π_i:

$$logit(\pi_i) = log \frac{\pi_i}{1-\pi_i} \in (-\infty, \infty)$$

  - This is the so-called "logit" link!
  - $\pi_i = E[Y_i|X_i]$: probability of having AD for a subject with covariates X_i.
  - $\frac{\pi_i}{1-\pi_i}$: odds
  - $logit(\pi_i) = log \frac{\pi_i}{1-\pi_i} = log \frac{P(Y_i=1|X_i)}{P(Y_i=0|X_i)}$: log-odds!

Logistic Regression

- We connect π_i and a linear function of the covariates X_i by assuming

$$log\frac{\pi_i}{1-\pi_i} = \beta_0 + x_{i1} \times \beta_1 + \ldots + x_{ip} \times \beta_p$$

- Essentially, we model the log-odds.
- But Y_i is a random variable. We need a distribution. A natural choice is the Bernoulli distribution

$$Y_i|X_i \sim Bernoulli(\pi_i)$$

- pmf, mean, variance:

Logistic Regression

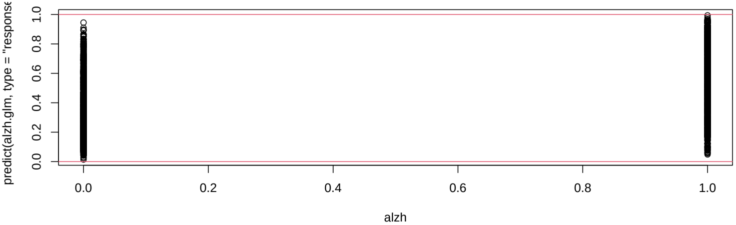
- Estimation of is typically conducted by maximizing the corresponding likelihood function
- How to obtain the likelihood function
  - $E(Y_i|X_i) = \pi_i = \frac{exp\{\beta_0 + x_{i1} \times \beta_1 + ... + x_{ip} \times \beta_p\}}{1 + exp\{\beta_0 + x_{i1} \times \beta_1 + ... + x_{ip} \times \beta_p\}}$
  - $f(Y_i|X_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$, i.e.,
    - $f(Y_i|X_i) = \pi_i$ if $Y_i = 1$
    - $f(Y_i|X_i) = 1 - \pi_i$ if $Y_i = 0$
  - independence: $f(Y|X) = \prod_{i=1}^n f(Y_i|X_i)$
  - $L(\beta_0, \beta_1, \cdots, \beta_p) = f(Y|X)$

Logistic Regression

- How to obtain the maximum likelihood estimates (MLE) of the parameters $(\beta_0, \cdots, \beta_p)$?
  - Iteratively re-weighted least squares (IRLS): the default method used by R
  - The Newton-Raphson algorithm

the Motivating Example of Logistic Regression

```
alzh.glm = glm(alzh ~ age + female + educ+lhippo, family=bino
par(mar = c(4, 4, 0.5, 0.5))
plot(alzh, predict(alzh.glm, type="response")); abline(h=c(0,
```




```
#More visualizaitons
#https://blogs.uoregon.edu/rclub/2016/04/05/plotting-your-log
```

Interpreting a logistic regression


```
summary(alzh.glm)$coefficients[-1,]
```

```
##            Estimate  Std. Error    z value     Pr(>|z|)
## age      0.01813761 0.004246088   4.271605 1.940715e-05
## female  -1.32020475 0.096534651 -13.675968 1.413151e-42
## educ    -0.05640342 0.013279326  -4.247461 2.162067e-05
## lhippo  -1.98502114 0.114028821 -17.408065 7.166544e-68
```

Interpreting a logistic regression

- Consider the age variable. The estimated coefficient is 0.018138. What information does it provide?
- The estimated log-odds AD for subject i is (or add a constant determined by study design, see Day 11 lecture)

$$loigt(\hat{\pi}_i) = \hat{\beta}_0 + \hat{\beta}_{age} age_i + \hat{\beta}_2 female_i + \hat{\beta}_3 educ_i + \hat{\beta}_4 lhippo_i$$

- Let $\tilde{\pi}_i$ denote estimated log-odds after one year

$$loigt(\tilde{\pi}_i) = \hat{\beta}_0 + \hat{\beta}_{age}(age_i + 1) + \hat{\beta}_2 female_i + \hat{\beta}_3 educ_i + \hat{\beta}_4 lhippo_i$$

Interpreting a logistic regression

- The estimated change in log-odds

$$logit(\tilde{\pi}_i) - logit(\hat{\pi}_i) = log\frac{\tilde{\pi}_i}{1-\tilde{\pi}_i} - log\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = 0.018138$$

- The odds of AD in one year later is $exp(0.018138) =$1.018303 times of the current odds.
- The estimated increase in odds of AD in a year is $e^{0.018138} - 1 = 1.8303\%$
- A 95% confidence interval
  - First, obtain a 95% C.I. for the difference in log-odds: $(0.018138 - 1.96 * 0.004246, 0.018138 + 1.96 * 0.004246) = (0.00982, 0.0265)$
  - Then, we transform them to increase in odds: $(e^{0.00982} - 1, e^{0.0265} - 1) = (0.99\%, 2.69\%)$

Interpreting a logistic regression

- What if we are interested in the increase in odds of AD in ten years (everything else is fixed)?
- The estimated increase in odds of AD in 10 years is

$$e^{10*0.018138} - 1 = 19.89\%$$

- A 95% C.I. for 10-year increase in odds:

```
exp(10*c(0.018138-1.96*0.004246, 0.018138+1.96*0.004246))-1
```

```
## [1] 0.1031375 0.3029118
```

i.e., $(10.3\%, 30.3\%)$

Interpreting a logistic regression

- Very often, we also want to know the significance of a variable after adjusting for other important covariates?
- Does age show a significant effect after adjusting for gender, education, and)hippocampus volume?
- A test for $H_0 : \beta_{age} = 0$ using the Wald test (a type of large-sample test)

```
summary(alzh.glm)$coefficients["age",]
```

```
##     Estimate   Std. Error      z value     Pr(>|z|)
## 1.813761e-02 4.246088e-03 4.271605e+00 1.940715e-05
```

- Other tests, such as likelihood ratio test, can also be used

GLM

- Recall that we used the logit link in the logistic regression

$$g(\pi_i) = logit(\pi_i) = \frac{\pi_i}{1-\pi_i},$$

where $\pi_i = E(Y_i|X_i)$.

- How about LM? $g(\mu_i) = \mu_i$, i.e., LM uses the identity link
- Poisson $g(\lambda_i) = log(\lambda_i)$. $Y_i|X_i \sim Poisson(\lambda_i)$.

Poisson Regression

Poisson Regression: The Model

- Poisson regression is often used to mode count data
- Why count data are special?
  - Count data are non-negative
  - Count data take integer values
- Count data often violate the assumption of "constant variance"
  - Count data often follow a Poisson distribution
  - Consider $K \sim Poisson(\lambda)$. $E(K) = ?, Var(K) =, pmf?$

Poisson Regression: Motivating Example

- Neurons may fire selectively for particular types of stimuli
- To understand whether a neuron is a visual-selective neuron, 20 trials were run for each of the five image categories:
  - animal, fruit, kids, military, space
- In each trial, the number of spikes (the number of times that the neuron fired) within a 1-second window was recorded

Poisson Regression


```
library(tidyverse)
#https://www.ics.uci.edu/~zhaoxia/Data/chosen_neuron_data.csv
chosen_neuron_data <- read_csv(
  "https://www.ics.uci.edu/~zhaoxia/Data/chosen_neuron_data.c
chosen_neuron_data <- chosen_neuron_data[, c(2:4)]
dim(chosen_neuron_data)
```

```
## [1] 100   3
```

```
names(chosen_neuron_data)
```

```
## [1] "trial_number" "n_spikes"     "image_categ"
```

Poisson Regression


```
attach(chosen_neuron_data)
# Even split of image categories among trials
table(image_categ)
```

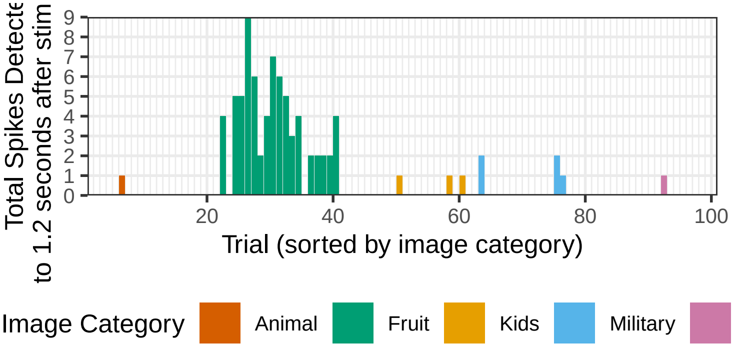
```
## image_categ
##    Animal     Fruit      Kids Military     Space
##        20        20        20        20        20
```

```
sapply(split(n_spikes, image_categ), mean)
```

```
##    Animal     Fruit      Kids Military     Space
##      0.05      3.60      0.15      0.25      0.05
```

Poisson Regression: Visualize the count data (by image category)


```
neuron_trial_plot
```



Poisson Regression


```r
# Fit generalized linear model
# Poisson GLM with the default log link function
poisson_fit <- glm(n_spikes ~ image_categ-1,
                   data = chosen_neuron_data,
                   family = poisson(link="log"))
# Tabulate the coefficient estimates
poisson_neuron_table <- summary(poisson_fit)$coefficients
row.names(poisson_neuron_table)=c("Animal", "Fruit",
                                  "Kids", "Military", "Space"
```

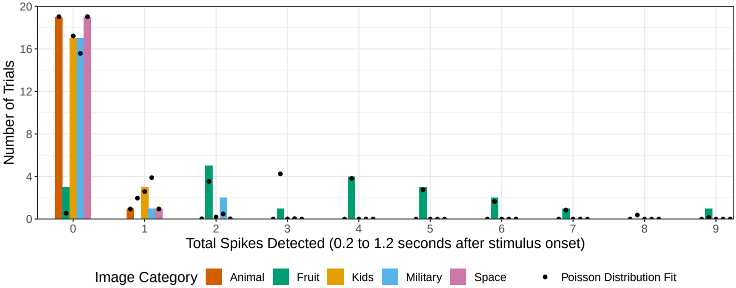
Poission Regression: Model Summary


```
poisson_neuron_table
```

```
##            Estimate Std. Error   z value     Pr(>|z|)
## Animal    -2.995732  0.9999998 -2.995733 2.737861e-03
## Fruit      1.280934  0.1178511 10.869084 1.618171e-27
## Kids      -1.897120  0.5773503 -3.285908 1.016541e-03
## Military -1.386294  0.4472132 -3.099851 1.936181e-03
## Space     -2.995732  0.9999998 -2.995733 2.737861e-03
```

Poisson Regression: Visualize Observed v.s. Fitted


```
poisson_obs_fit_plt
```



The Deviance of GLM object

- Next, we would like to discuss the significance of the image_categ variable. To do so, we first look at the deviance of a GLM object
- The deviance of a GLM object obj is

$$2[log(L_{saturated}) - log(L_{obj})]$$

- What is the saturated model?
  - Logistic: $\pi_i = y_i$ and $L_{saturated} = 1$
  - Poisson: $\lambda_i = y_i$ and $L_{saturated} = \prod_i \frac{y_i^{y_i} e^{-y_i}}{y_i!}$.

Poission Regression: The Overall Significance

- Consider two <u>nested</u> models obj1 and obj2, the difference in their deviances is

$$2[log(L_{obj2}) - log(L_{obj1})],$$

which is the LRT statistic.

Poission Regression: The Overall Significance

```
# Test for visual selectivity: Likelihood Ratio Test
poisson_fit0 = glm(n_spikes ~1, data=chosen_neuron_data, fami
anova(poisson_fit0, poisson_fit, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: n_spikes ~ 1
## Model 2: n_spikes ~ image_categ - 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        99    260.985
## 2        95     81.213  4   179.77 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Poission Regression: The Overall Significance

```
# Test for visual selectivity: Rao's score test
anova(poisson_fit0, poisson_fit, test = "Rao")
```

```
## Analysis of Deviance Table
##
## Model 1: n_spikes ~ 1
## Model 2: n_spikes ~ image_categ - 1
##   Resid. Df Resid. Dev Df Deviance   Rao  Pr(>Chi)
## 1        99    260.985
## 2        95     81.213  4   179.77 236.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Poission Regression: The Overall Significance


```
# Test for visual selectivity: Wald test
Wald.stat=poisson_fit$coefficients %*%
  solve (summary(poisson_fit)$cov.unscaled) %*%
  poisson_fit$coefficients
1-pchisq(Wald.stat, df=4)
```

```
##      [,1]
## [1,]    0
```

Poission Regression: Model Interpretation


```
## Estimate Std. Error z value Pr(>|z|)
## Animal -2.995732 0.9999998 -2.995733 2.737861e-03
```

- $\hat{\beta}_{Fruit} = 1.2809$: What does it tell us?
- Recall that we used the log link

Poission Regression: Model Interpretation

- Note that the model poisson_fit does not include β_0.
- That's why we can estimate β_{Animal}, β_{Fruit}, β_{Kids}, $\beta_{Military}$, β_{space}.
- Question: how should we interprete the estimated coefficients if the intercept term was included?
  - Try poisson_fit_repara <- glm(n_spikes ~ image_categ, data = chosen_neuron_data, family = poisson(link="log"))
  - Are the two models equivalent? (Lab activity)

Poisson Regression: Model Interpretation

- Re-parameterization
- Parameters:
  - poisson_fit:
  - poisson_fit_repara:

Linear Functions of Parameters

- The Poisson regression we fit provides estimates of β_{Animal}, β_{Fruit}, β_{Kids}, $\beta_{Military}$, β_{space}, which are the log of the Poisson rates
- What if we are interested in difference between specific groups? e.g.,
  - $\beta_{Fruit} - \beta_{Animal}$
  - $\frac{\beta_{Fruit}+\beta_{Animal}+\beta_{Kids}+\beta_{Military}+\beta_{Space}}{5}$
  - $\beta_{Fruit} - \frac{\beta_{Animal}+\beta_{Kids}+\beta_{Military}+\beta_{Space}}{4}$
- They are linear functions of the coefficients, i.e., in the form of $a^T\beta$, where a is a 5-by-1 vector.

Linear Functions of Parameters

- LM/GLM provides not only estimated coefficients but also the variance-covariance of the estimated covariates
  - Let $\hat{\beta} = c(\hat{\beta}_1, \cdots, \hat{\beta}_p)^T$
  - Let $\hat{\Sigma}$ denote the estimated variance-covariance of $\hat{\beta}$
  - Let a be linear coefficients

Inference of Linear Functions of Parameters

- Consider a linear function : $a^T\beta$
- Estimate: $a^T\hat{\beta}$
- Variance of the estimate: $Var(a^T\hat{\beta}) = a^T\hat{\Sigma}a$
- Standard Error (SE): $s.e.(a^T\hat{\beta}) = \sqrt{a^T\hat{\Sigma}a}$
- A 95% confidence interval:

$$(a^T\hat{\beta} - 1.96 * s.e.(a^T\hat{\beta}), a^T\hat{\beta} + 1.96 * s.e.(a^T\hat{\beta}))$$

- Z-value: $\frac{a^T\hat{\beta} - ?????}{s.e.(a^T\hat{\beta})}$

Inference of Linear Functions of Parameters: Example

- Parameter of interest: $\frac{\beta_{Fruit}+\beta_{Animal}+\beta_{Kids}+\beta_{Military}+\beta_{Space}}{5}$

```
#extract estimated coefficients and their var-cov
poisson_fit$coefficients
```

```
##   image_categAnimal    image_categFruit     image_categKid
##           -2.995732            1.280934           -1.89712
##    image_categSpace
##           -2.995732
```

```
summary(poisson_fit)$cov.unscaled #this ia a 5-by5 matrix
```

```
##                       image_categAnimal image_categFruit ima
## image_categAnimal             0.9999996       0.00000000
## image_categFruit              0.0000000       0.01388889
```

Inference of Linear Functions of Parameters: Example

```
a=matrix(rep(1/5,5), 1)
a%*%poisson_fit$coefficients #estimate
```

```
##              [,1]
## [1,] -1.598789
```

```
sqrt(a%*%summary(poisson_fit)$cov.unscaled%*%t(a)) #s.e.
```

```
##              [,1]
## [1,] 0.3192003
```

Linear Contrasts

- Linear contrasts are a special family of linear functions
- We say $a^T\beta = \sum_i a_i\beta_i$ is a linear contrast if $\sum a_i = 0$, where $a = (a_1, \cdots, a_p)^T$.
- Often, we are interested in whether a linear contrast is zero, i.e., $H_0 : a^T\beta = 0$
- z-value: $\frac{a^T\hat{\beta}-0}{s.e.(a^T\hat{\beta})}$

Linear Contrast (e.g., Fruit vs Animal)


```
a <- matrix(c(-1, 1, 0, 0, 0), 1)
#estimate
fruit_animal_est = a%*%poisson_fit$coefficients
#variance
fruit_animal_var = a%*%summary(poisson_fit)$cov.unscaled%*%t(
#z value
print(fruit_animal_est/sqrt(fruit_animal_var))
```

```
##           [,1]
## [1,] 4.247274
```

Linear Contrast (e.g., Fruit vs Animal) with the multcomp package


```
library(multcomp)
a <- matrix(c(-1, 1, 0, 0, 0), 1)
t <- glht(poisson_fit, linfct = a)
summary(t)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = n_spikes ~ image_categ - 1, family = po
##     data = chosen_neuron_data)
##
## Linear Hypotheses:
##         Estimate Std. Error z value Pr(>|z|)
## 1 == 0     4.277      1.007   4.247 2.16e-05 ***
```

Multinomail Logistic Regression


```
library(nnet)
multinom(diagnosis ~ age + female + educ + lhippo + rhippo)
```

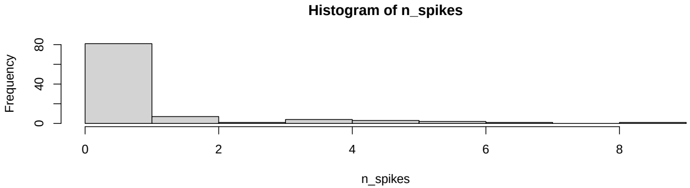
```
## # weights:  21 (12 variable)
## initial  value 2966.253179
## iter  10 value 2372.326777
## final  value 2288.461323
## converged
```

```
## Call:
## multinom(formula = diagnosis ~ age + female + educ + lhipp
##     rhippo)
##
## Coefficients:
##   (Intercept)         age    female        educ    lhippo
## 1    3.671844 0.026773068 -1.237127 -0.04694669 -1.251316
```

Other concerns

- Dispersion: under- or over-dispersion
- Zero-inflated Poisson Regression
- Model selection ...

```
hist(n_spikes)
```




```
#Interested inhow to fit a zero-inflated Poisson regression?
#https://www.rdocumentation.org/packages/pscl/versions/1.5.5
```

The Assumption of Independence

Independent Observations

- The common assumption we have made in LM and GLM is that the observations are independent with each other
- This is not always the case
- Examples:

Independent Observations

- What is the consequence of ignoring data independence?
  - The damage is probably worse than violations of distributions
  - Fortunately, tools have been developed to account for data dependence
  - Day 13: Linear Mixed-Effects Model (LME)
  - Day 14: Generalized Linear Mixed-Effects Model (GLMM)