ST 437/537: Applied Multivariate and Longitudinal Data Analysis

Longitudinal Data Analysis: Review of Generalized Linear Model

Arnab Maity
NCSU Department of Statistics
SAS Hall 5240 919-515-1937 amaity[at]ncsu.edu

References:

- Applied Regression Modeling by Iain Pardoe. Wiley.
- Modeling Longitudinal Data by Robert E. Weiss. New York: Springer.
- Linear Mixed Models for Longitudinal Data by Geert Verbeke and Geert Molenberghs. New York: Springer.
- Applied Longitudinal Analysis by Fitzmaurice by G.M., Laird, N.M., and Ware, J.H. New York: Wiley (on reserve at NCSU library)

Introduction

In the previous chapters we focused on methods for analyzing longitudinal data where

- the response variable is continuous with values ranging over the real line;
- the vector of subject-level responses is assumed to have (exactly or approximately) a multivariate normal distribution.

In this chapter, we consider the case where the vector of subject-level **responses cannot be modeled using a normal distribution**.

Some examples of this type of data:

- the response is binary and takes only values 0(“failure”) ot 1 (“success”)
- the response is a “count” $(0, 1, 2, \ldots)$, but the values are relatively small, etc.

We refer to these types of responses by the name **generalized responses**. The models used to analyze generalized responses, analogous to linear models, are called **generalized linear models**. We begin with a review of the generalized linear models for scalar responses and then discuss these class of models for repeated measures.

Binary response: Logistic Regression

A popular regression model for binary response is the logistic regression model. Let us start with an example.

Example: [Applied Regression Modeling by Iain Pardoe]. Bennett and Mangasarian (1992) presented the data set from the Wisconsin Breast Cancer Study consisting of 683 cases of potentially cancerous tumors. Out of them, 444 turned out to be benign and 239 were malignant. Typically, an invasive surgical biopsy procedure is used to used to determine whether a tumor is malignant or benign. An alternative, less invasive technique called “fine needle aspiration” (FNA) allows examination of a small amount of tissue from the tumor. In our data set, FNA provided nine cell features for each case; a biopsy was then used to determine the tumor status as malignant or benign.

Question of interest: Is it possible to determine whether and how these cell features tell us anything about the tumor status?

Let us take a look at a snapshot of the data set.

```
dat <- as.matrix(read.csv("data/bcwis.csv", header = , sep = ","))
head(dat)
```

```
##           ID Y X1 X2 X3 X4 X5 X6 X7 X8 X9
## [1,] 1000025 0  5  1  1  1  2  1  3  1  1
## [2,] 1002945 0  5  4  4  5  7 10  3  2  1
## [3,] 1015425 0  3  1  1  1  2  2  3  1  1
## [4,] 1016277 0  6  8  8  1  3  4  3  7  1
## [5,] 1017023 0  4  1  1  3  2  1  3  1  1
## [6,] 1017122 1  8 10 10  8  7 10  9  7  1
```

In our case Y is the response, and Y can take only two values: 1 (if malignant) and 0 (benign). The other variables X_1 to X_9 are predictors.

> *Since Y is a binary variable, we can not directly use a linear regression model.*

Instead we use the logistic regression. Specifically, we assume

$$Y_i \sim Binomial(1, p_i),$$

where $p_i = P(Y_i = 1)$. We **model the probability** $p_i = P(Y_i = 1)$ as function of the covariates:

$$p_i = P(Y_i = 1) = \frac{exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik})}{1 + exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik})}.$$

Another way to write the same model is

$$log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}.$$

The transformation in the left hand side is called the *logit* transformation.

To estimate the regression parameters $\beta_0, \ldots, \beta_k$, we use maximum likelihood estimator (since we know the distribution of Y_i, we can write the likelihhod function). The is no closed form expression for $\widehat{\beta}_0, \ldots, \widehat{\beta}_k$, and a software must be used to obtain them.

In R, we can use `glm` function in R. Let us fit the model for our data set.

```
## Define the covariates
X <- dat[, -c(1,2)]

## Define the response
Y <- dat[, 2]

## Fit the logistic regression
out <- glm(Y ~ X, family = binomial())
```

The first argument in the `glm` function is the usual formula `Y ~ X` to indicate that Y should be regressed onto X. The second argument `family = binomial()` specifies the distribution of Y (recall we assumed Y follows a binomial distribution).

```
summary(out)
```

```
##
## Call:
## glm(formula = Y ~ X, family = binomial())
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4841  -0.1153  -0.0619   0.0222   2.4698
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.10394    1.17488  -8.600  < 2e-16 ***
## XX1           0.53501    0.14202   3.767 0.000165 ***
## XX2          -0.00628    0.20908  -0.030 0.976039
## XX3           0.32271    0.23060   1.399 0.161688
## XX4           0.33064    0.12345   2.678 0.007400 **
## XX5           0.09663    0.15659   0.617 0.537159
## XX6           0.38303    0.09384   4.082 4.47e-05 ***
## XX7           0.44719    0.17138   2.609 0.009073 **
## XX8           0.21303    0.11287   1.887 0.059115 .
## XX9           0.53484    0.32877   1.627 0.103788
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 102.89  on 673  degrees of freedom
## AIC: 122.89
##
## Number of Fisher Scoring iterations: 8
```

The output is very similar to output from a standard linear regression. The main differences are as follows.

- The test statistics for each β coefficients is the z-statistics (not the t-statistics). This is a large sample approximation.
- The is no residual standard error (i.e., no estimate of σ^2)

Testing hypothesis: Notice that X_2 and X_5 have high p-values. Thus we can try to perform a formal test for the hypothesis

$$H_0 : \beta_2 = \beta_5 = 0$$

versus at least one of β_2 or β_5 not zero. We can still use the **likelihood ratio test** using the `deviance` in the output. The test statistic is the difference in the deviances of the full model fit and the reduced model fit. Under H_0, this statistic follows a χ^2_d distribution where d is the difference in the number of parameters between the two models.

```
## Full model fit
full.model <- out

## Reduced model fit
newX <- X[, -c(2,5)]
reduced.model <- glm(Y ~ newX, family = binomial())

## Deviances
anova(reduced.model, full.model)
```

```
## Analysis of Deviance Table
##
## Model 1: Y ~ newX
## Model 2: Y ~ X
##   Resid. Df Resid. Dev Df Deviance
## 1       675     103.27
## 2       673     102.89  2  0.37857
```

```
## Computation of p-value
pchisq(0.37857, df=2, lower.tail=F)
```

```
## [1] 0.8275506
```

The high p-value indicates that we can not reject $H_0 : \beta_2 = \beta_5 = 0$. The we can omit these two variables from our model.

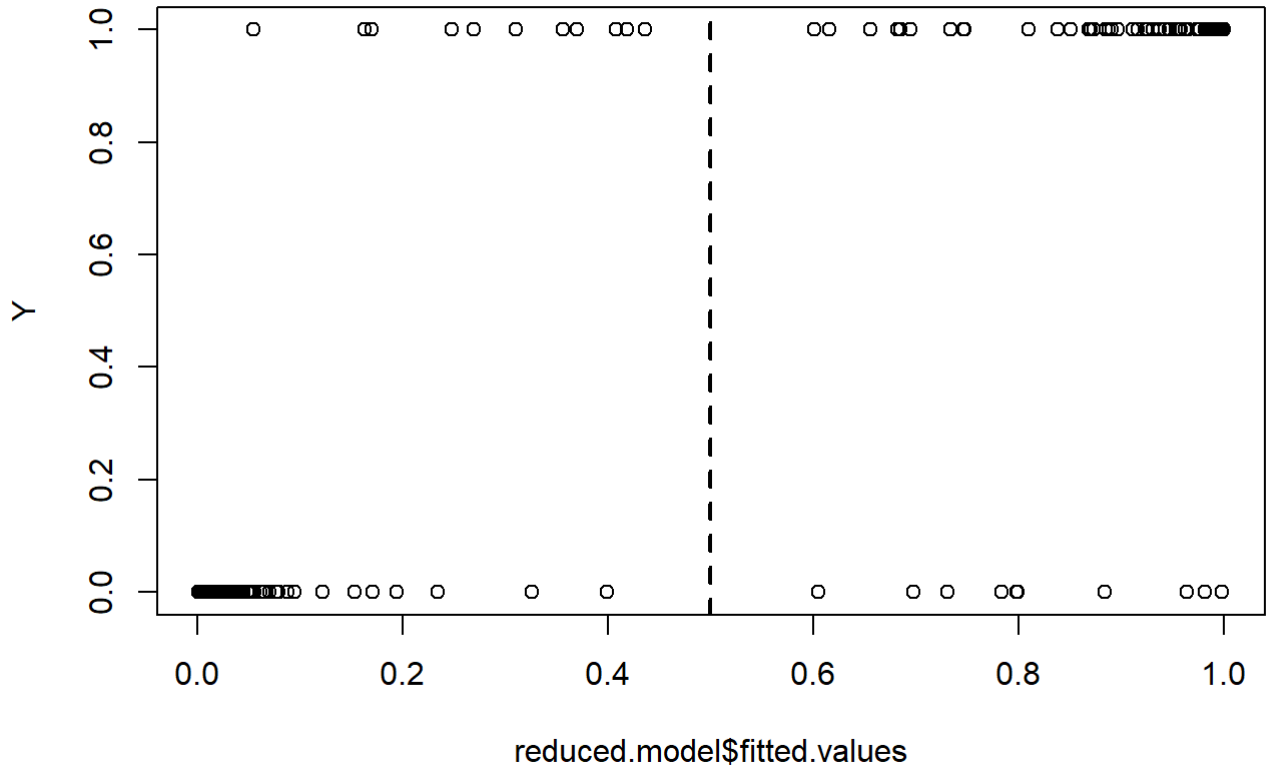
Prediction: We can find the predicted probabilities from the output as follows.

```
head(out$fitted.values)
```

```
##           1           2           3           4           5           6
## 0.016046581 0.908808622 0.008137623 0.760934919 0.018166848 0.999973622
```

If a probability for a particular individual $P(Y_i = 1)$ is large (closer to 1), we would predict $\widehat{Y}_i$ to be 1; 0 otherwise.

```
plot(reduced.model$fitted.values, Y)
abline(v = 0.5, lty=2, lwd=2)
```



The vertical line is drawn at probability 0.5. Based on this, we can classify all tumors to the left of this cutoff (< 0.5) as benign ($Y = 0$) and all tumors to the right of this cutoff (> 0.5) as malignant ($Y = 1$).

Count data: Poisson regression

Poisson regression is typically used (among a few available other models) when the response variable is a count.

Example: [Applied Regression Modeling by Iain Pardoe]. “Durham et al. (2004) analyzed wine demand at a restaurant, using economic hedonic quantity models to evaluate the impact of objective factors (e.g., origin, varietal), sensory descriptors, and price, on the choice of restaurant wines. The data were collected at a high-end restaurant over 19 weeks in 1998.”

In this situation, the **predictors are the wine features**, and the **outcome Y is a count** of the number of bottles of each wine sold in each of the 19 weeks of the study. Thus the only possible values of Y that is, the only possible values are nonnegative integers: 0, 1, 2, .

Poisson regression assumes

$$Y_i \sim Poisson(\lambda_i),$$

where $\lambda_i = E(Y_i)$ and the the expected value of Y_i, $E(Y_i)$, as a function of covariates in the *log-linear* fashion:

$$\lambda_i = log[E(Y_i)] = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}.$$

The `log` transformation ensures that the mean count $E(Y_i)$ is always non-negative.

We can fit this model using `glm`.

```
dat <- as.matrix(read.csv("data/winewhite.csv", header = , sep = ","))
head(dat)
```

```
##      ID  Y  X D1 D2 D3 D4 D5 D6 D7 D8 D9 D10 D11 D12 D13 D14 D15 D16 D17
## [1,] 47 15 19  0  1  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0
## [2,] 47 17 19  0  1  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0
## [3,] 47  6 19  0  1  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0
## [4,] 47 12 19  0  1  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0
## [5,] 47  9 19  0  1  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0
## [6,] 47 11 19  0  1  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0
##      D18 D19 D20
## [1,]   0   0   0
## [2,]   0   0   0
## [3,]   0   0   0
## [4,]   0   0   0
## [5,]   0   0   0
## [6,]   0   0   0
```

```
## Covariates
D <- dat[, -c(1, 2, 3)]
X <- dat[, 3]

## Response
Y <- dat[, 2]

## glm fit
out <- glm(Y ~ X + D, family = poisson())
```

The second argument `family = poisson()` specifies the distribution of Y (recall we assumed Y follows a poisson distribution).

```
summary(out)
```

```
##
## Call:
## glm(formula = Y ~ X + D, family = poisson())
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0383  -0.7851  -0.3908   0.0962   3.3255
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.72639    0.69085   3.946 7.93e-05 ***
## X           -0.10514    0.02898  -3.628 0.000285 ***
## DD1         -1.14591    0.37282  -3.074 0.002115 **
## DD2          1.63993    0.23199   7.069 1.56e-12 ***
## DD3         -1.38458    0.16627  -8.327  < 2e-16 ***
## DD4         -0.83931    0.12752  -6.582 4.65e-11 ***
## DD5         -1.04371    0.17686  -5.901 3.60e-09 ***
## DD6         -1.02050    0.19133  -5.334 9.62e-08 ***
## DD7         -1.86373    0.53222  -3.502 0.000462 ***
## DD8         -0.71086    0.38288  -1.857 0.063364 .
## DD9         -0.18541    0.52048  -0.356 0.721662
## DD10         1.01609    0.38483   2.640 0.008281 **
## DD11         0.45177    0.36025   1.254 0.209826
## DD12        -0.11134    0.32808  -0.339 0.734338
## DD13         0.25282    0.42108   0.600 0.548229
## DD14         1.21592    0.41211   2.950 0.003173 **
## DD15         0.58429    0.23230   2.515 0.011895 *
## DD16         0.09414    0.30596   0.308 0.758308
## DD17        -0.11398    0.33488  -0.340 0.733584
## DD18        -0.82829    0.36236  -2.286 0.022266 *
## DD19         0.24458    0.40023   0.611 0.541125
## DD20        -0.14155    0.28748  -0.492 0.622438
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2046.92  on 550  degrees of freedom
## Residual deviance:  492.17  on 529  degrees of freedom
## AIC: 1123.1
##
## Number of Fisher Scoring iterations: 6
```

We can test for hypothesis using likelihood ratio test as well.

Generalized Linear Model

Generalized linear models extend the methods of regression analysis to settings where the outcome is dichotomous (binary) variable, count etc. They share many of the characteristics of linear models; most notably the fact that a linear combination of the covariates is related to the mean response. They differ from the linear model in couple of ways including the fact that the distribution of the response may not be normal. Instead the distribution of the response is assumed to be in the **exponential family** of distributions. The exponential family class is very wide and includes normal distribution, but also Binomial, Poisson, Gamma and many more distributions.

In general, observed data are: $[Y_i, X_{i1}, \ldots, X_{iK}]$ for $i = 1, \ldots, n$ where Y_i is the scalar response and $X_{i1}, \ldots, X_{iK}$ are covariates. It is assumed that Y_i is in the family of exponential family models, and Y_i's are independent over i. The **generalized linear models** are specified by the following **THREE main parts**:

- **distributional assumption** of Y_i
- modeling how the covariates affect the mean response (**the systematic component**)
- specification of how the the mean response is linked to the systematic component (**link function**).

The distributional assumption

It is assumed that the outcome Y_i is modeled using a distribution that belongs to the exponential family. Examples:

- $Y_i \sim Bernoulli(p_i)$. This means

$$P(Y_i = y) = p_i^y (1 - p_i)^{1-y}.$$

  What are the mean and variance?

- $Y_i \sim Poisson(\lambda_i)$. This means

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

  What are the mean and variance?

- $Y_i \sim N(\mu_i, \sigma^2)$. This means

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu_i)^2}$$

What are the mean and variance?

The systematic component

The systematic component specifies that the effect of the covariates $X_{i1}, \ldots, X_{iK}$ on the mean response Y_i can be expressed in terms of the following linear predictor

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_K X_{iK}$$

or in vector format $\eta_i = X_i^T \beta$, where X_i is the $(K+1)$-dimensional column vector of X_{il}'s with 1 as the first element, and $\beta = (\beta_0, \beta_1, \ldots, \beta_K)^T$ is the $(K+1)$-dimensional column vector of β_l's. The parameter η_i is called the . The parameter β is called **regression parameter**.

The link function

The *link function* applies a transformation to the mean response and then links the transformed mean to the covariates (through the linear predictor).

Denote the mean response $\mu_i = E[Y_i]$; then the common notation for the link function is: $g(\mu_i) = \eta_i$.

The link function is known and assumed monotone and differentiable over the domain of μ_i.

Examples:

- Identity link (common for normal responses), $g(x) = x$:

- Logit link (common for binary responses $0/1$, logistic regression), $g(p) = log\frac{p}{1-p}$
- Probit link (used for binary responses, probit regression) $g(x) = \Phi^{-1}(x)$ where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal variable $N(0, 1)$.
- Log link (used for counts responses, e.g., poisson regression) $g(x) = log(x)$

Combining all the three parts, we have

$$Y_i \sim EF, \qquad g\left(E[Y_i]\right) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_K X_{iK}$$

- Logistic regression model: $Y_i \sim Bernoulli(p_i)$ with $log\{p_i/(1-p_i)\} = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_K X_{iK}$
- Poisson regression model: $Y_i \sim Poisson(\mu_i)$ with $\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_K X_{iK}$

Note: In the Poisson regression model above, we assume that the response is calculated over the same duration (i.e., number of tonadoes per year, number of telephone calles per 10 minutes etc.).

If we measure the i-th response Y_i over a time period of length T_i, then the model becomes

$$Y_i \sim Poisson(\mu_i) \quad \text{with} \quad \log\left(\frac{\mu_i}{T_i}\right) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_K X_{iK}.$$

Parameter estimation

As GLMs specify a distribution for the response Y_i, maximum likelihood estimation (MLE) is used to estimate the model parameters.

- The estimator $\widehat{\beta}$ of β does not often have any closed form.
- For large sample size n, we can show $\widehat{\beta}$ is approximately normal and unbiased.
- For large sample size, we can use the z-test to test individual parameters (as seen in the outputs from `glm`)

Main page: **ST 437/537: Applied Multivariate and Longitudinal Data Analysis (https://maityst537.wordpress.ncsu.edu/)**