GLMs: Generalized Linear Models

Professor Diane Lambert

June 2010

Supported by MOE-Microsoft Key Laboratory of Statistics and Information Technology and the Beijing International Center for Mathematical Research, Peking University.

With many thanks to Professor Bin Yu of University of California Berkeley, and Professor Yan Yao and Professor Ming Jiang of Peking University.

Linear Regression Models

The mean is linear in **X**

$$E(Y \mid \mathbf{X}) = \mu(\mathbf{X}) = \mathbf{X}\beta = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K$$

The variance is constant in **X**

$$\mathrm{var}(Y \mid \mathbf{X}) = \sigma^2$$

Y doesn't have to be normal (just use the CLT), but it should have more than a few values.

These assumptions can be unreasonable.

Linear Regression & The Poisson

$Y \mid \mathbf{X} \sim$ Poisson with mean $\mu(\mathbf{X})$

a) $\mathrm{var}(Y \mid \mathbf{X}) = \sigma^2(\mathbf{X}) = \mu(\mathbf{X})$,

which isn't constant

b) the mean is positive,

often $\mu(\mathbf{X})$ is not linear

instead effects multiply instead of add

$\mu(\mathbf{X}) = \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K$

Modeling log(Y) doesn't help

$\log(0) = -\infty$

$\mathrm{var}(\log(Y) \mid \mathbf{X}) \approx 1/\mu(\mathbf{X})$, which isn't constant

Linear Regression & Binary Data

$$Y \mid \mathbf{X} \sim \begin{cases} 1 & \textit{with probability} \quad \mu(\mathbf{X}) \\ 0 & \textit{with probability} \quad 1-\mu(\mathbf{X}) \end{cases}$$

a) $\sigma^2(\mathbf{X}) = \mu(\mathbf{X})(1- \mu(\mathbf{X})) \neq$ constant

b) $0 \leq m(X) \leq 1$

c) linear differences in $\mu(\mathbf{X})$ aren't important

changing from .10 to .01 or .9 to .99 is more extreme than changing from .6 to .51 or .69

Transforming Y doesn't help

Y will still have only two values

Generalized Linear Models (GLMs)

1. The mean outcome $\mu(\mathbf{X})$ of Y is connected to a linear combination of $\mathbf{X}$ by a link function g

$$g(\mu(\mathbf{X})) = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K$$

2. $\sigma^2(\mathbf{X})$ can depend on $\mu(\mathbf{X})$

$$\sigma^2(X) = V(\mu(\mathbf{X}))$$

Transforming the mean (not the outcome) to get linearity.

Examples

linear regression: g = I, V is constant

log-linear (Poisson) regression: g = log, V = I

Logistic Regression

It's a GLM

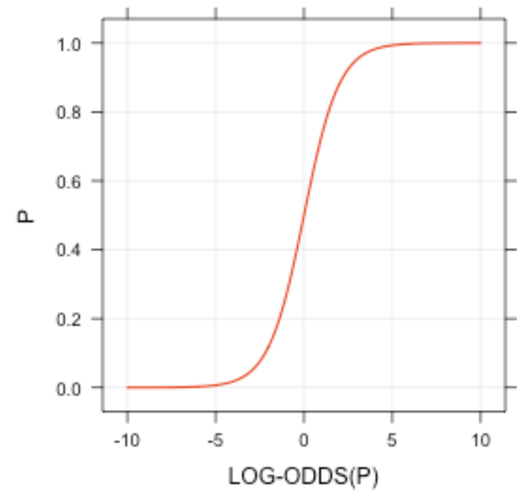
Y is binary with mean $\mu(\mathbf{X})$

link: $g(\mu) = \log(\mu/(1- \mu)) = \mathrm{logit}(\mu)$

$g(\mu)$, the log odds, is linear

stretches small and large μ

var: $\sigma^2(X) = \mu(\mathbf{X})(1 - \mu(\mathbf{X}))$

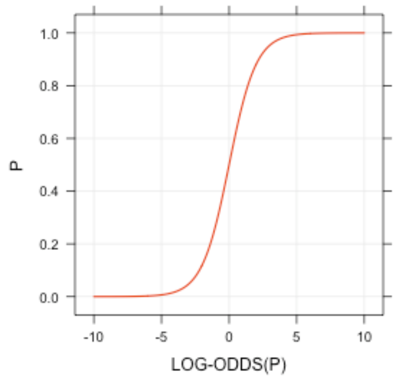


Any model with this link and variance function could be called logistic regression, but the term is usually reserved for binary data

use `qlogis` in R to compute logit(p) = log-odds(p).

The Logit Link Function

The intercept in logistic regression does not shift the mean by a constant.



logit(μ) = log(μ/(1- μ)) = β_0

Increasing β_0 by .4 increases μ by

.1 at μ = .5 since logit(.5) = 0, logit(.6) = .4

.06 at μ = .8

.03 at μ = .9

.003 at μ = .99

Effects are linear on the log-odds scale but smaller in the tails on the *probability* scale.

Logistic Regression Coefficients

Some people like to interpret logistic regression coefficients on the odds scale

$\text{odds}(\mu) = \mu/(1-\mu) = P(Y=1)/P(Y=0)$

$\log(\text{odds}(\mu)) = \text{logit}(\mu)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K$

Increasing X_1 by 1

adds β_1 to $\log(\text{odds}(\mu))$

multiplies the odds of a success by $\exp(\beta_1)$

Another GLM for Binary Outcomes

Probit Regression

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

Φ^{-1} gives quantiles for a normal(0,1)

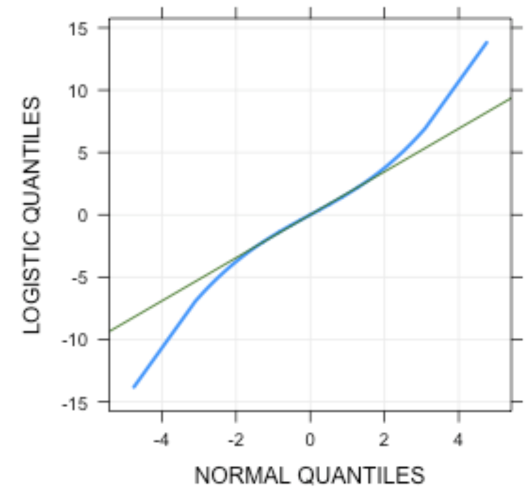
log-odds gives quantiles for a logistic

$\Phi^{-1}(p) \approx$ log-odds(p) over (.05, .95)

logistic is more extreme beyond

when $\Phi^{-1}(p) = -4$, log-odds(p) = -7

probit regression is popular in economics and sometimes Bayesian modeling



green line: regression of the logistic quantiles on the normal over [-2, 2]

Fitting GLMs in R

logistic regression

```
z <- glm(formula, family = binomial)
```

probit regression

```
z <- glm(formula,
         family=binomial(link='probit'))
```

log-linear regression

```
z <- glm(formula, family = Poisson)
```

GLM coefficients

are MLEs

computed iteratively, weighted least squares at each step.

Weighted Least Squares

Ordinary least squares estimate

$$b = (X'X)^{-1}X'Y$$

Each (X_i, Y_i) is treated the same

Agrees with the assumption of constant variance for $Y|X$ in linear regression.

Weighted Least Squares

Different Y's have different variances

$w_i = 1/\mathrm{var}(Y_i|X_i)$ $W = \mathrm{diag}(w)$

$$b = (X'WX)^{-1}X'WY$$

observations with big variances are downweighted

GLMs & Weighted Least Squares

Weighted Least Squares

Different Y's have different variances

$w_i = 1/var(Y_i|X_i)$ $W = diag(w)$

$$b = (X'WX)^{-1}X'WY$$

observations with big variances are downweighted

In a GLM, the $var(Y_i|X_i)$ depends on the unknown b.

Strategy

- Get an guess for μ_i (e.g., from ordinary least squares)
- Compute $W = diag(1/V(\mu))$ using the variance function
- Compute weighted least squares, get new μ, new W,
- update weighted least squares, etc.

Goodness of Fit

Linear Regression

$$R^2 = 1 - \frac{(n-K-1)^{-1}\sum\left(Y_i - b_0 - b_1 X_1 - \ldots - b_K X_K\right)^2}{(n-1)^{-1}\sum\left(Y_i - \overline{Y}\right)^2}$$

Compares residuals under the fitted model to those under the null (no predictors) model

R^2 is not sensible if var(Y|**X**) is not constant

- In that case, some Y’s are noiser than others, so we shouldn’t worry about their residuals as much

Deviance: Goodness of Fit for GLMs

Choose a probability family $p(y_i \mid \beta)$

binomial for logistic regression

Poisson for loglinear regression

loglik(β| $y_1, \ldots, y_n$) = $L(\beta|\mathbf{y}) = \sum_i \log(p(y_i| \beta)$

maximum likelihood estimates maximize log-likelihood

Deviance

$$D(\beta) = -2[L(\beta|\mathbf{y}) - L(\beta_S|\mathbf{y})]$$

β_S gives the saturated model with n parameters for n y_i's

For logistic regression

$$D(\beta) = -2\left(\sum_i y_i \log(\mu_i / y_i) + (1 - y_i)\log\left((1-\mu_i)/(1-y_i)\right)\right)$$

GLM

- Model of the mean
  - transform with link functions to linearity
  - in exponential families, this is often the natural parametrization
    - log for Poisson; logit for binomial
- Model of the variance
  - variance is a function of the mean
- Goodness of fit measure
  - deviance

Logistic Regression Example

Because of geography, many wells in Bangladesh are contaminated with arsenic.

Bangladesh assumes safe limit = 50 μg/l

World Health Organization assumes safe limit = 10 μg/l

Wells near unsafe ones can still be safe

3/4 of safe well owners would share drinking water

Owners of unsafe wells were advised to switch

Outcome: did owners of unsafe wells switch?

Predictors: what influenced the decision to switch

The Data

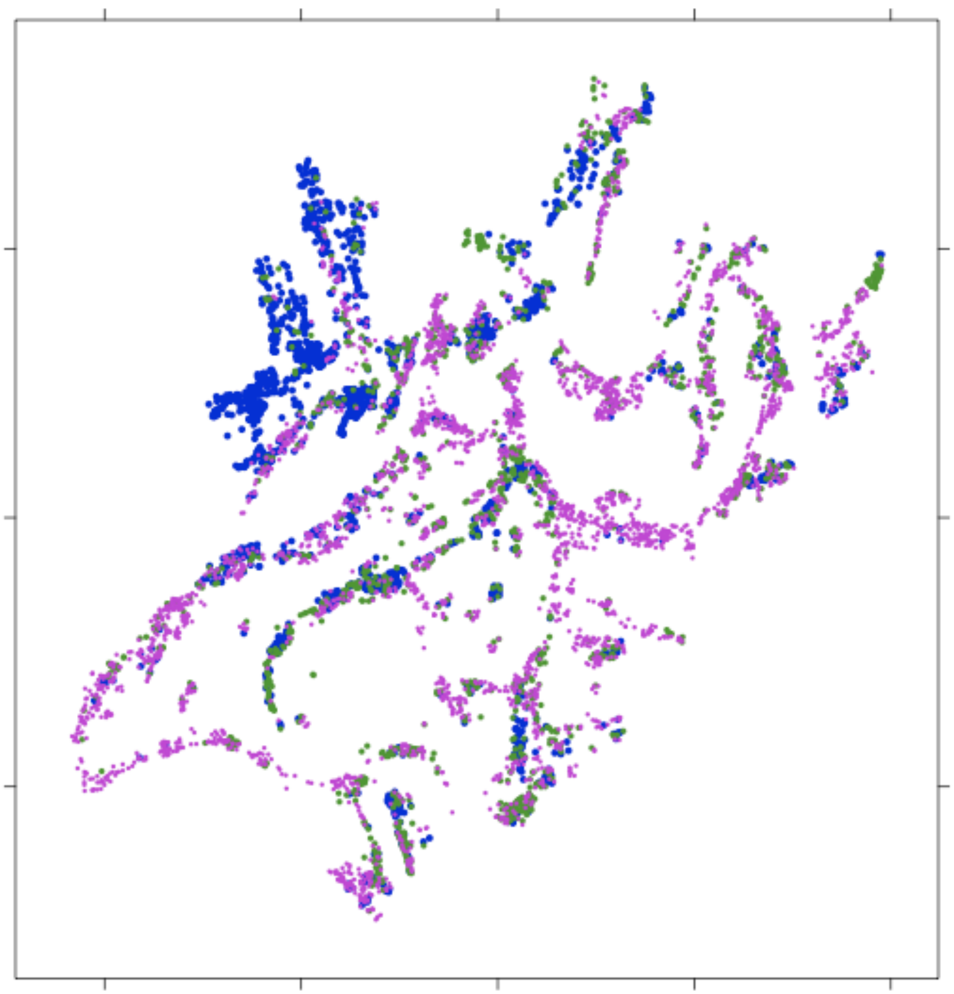
outcome: switch

predictors:

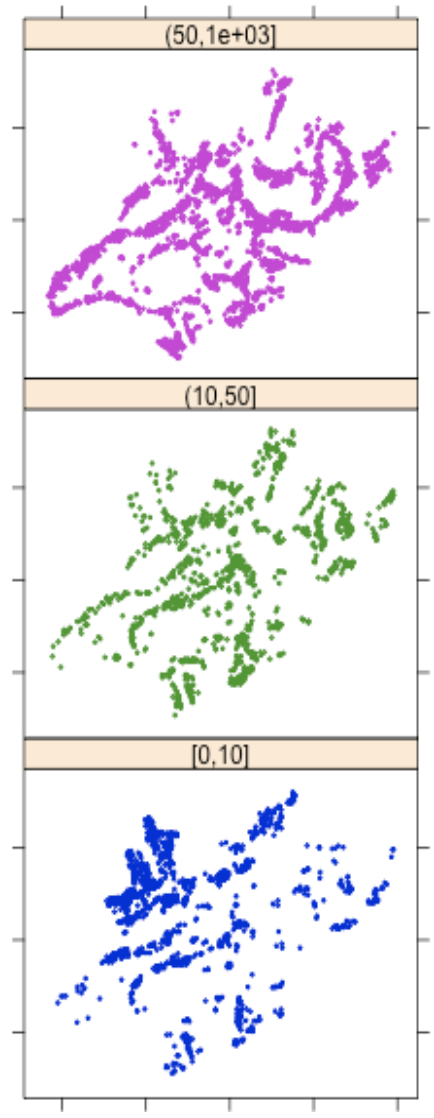
- arsenic
- unsafe
- distance
- ‘lat’
- ‘long’
- community
- education

3070 safe wells; 3378 unsafe wells

Locations of Wells



[0,10] μg/l (10, 50] >50



Logistic Regression

Much of what we learned about linear regression applies to logistic regression.

- think about the outcome
  - switching when the well is unsafe
- think about which variables matter most
  - arsenic level?
  - distance from the nearest safe well?
- think about scales
  - log distance? truncate?
- interactions?

Logistic Regression Example

Start by assuming people won't go more than 10 km to get drinking water

```
wells$walkDistance <
   pmin(wells$distance/1000, 10)

zArDist <- glm(switch ~ walkDistance +
                        log(arsenic),
               data = wells,
               subset = unsafe,
               family = binomial)
```

R Output

```
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.10165    0.30921 -10.031   <2e-16
walkDistance -0.12014    0.01305  -9.204   <2e-16
log(arsenic)  0.84454    0.06580  12.834   <2e-16

Null deviance:     4486.8  on 3377 deg of freedom
Residual deviance: 4269.2  on 3375 deg of freedom
```

Deviance is expected to decrease by 1 when an unnecessary predictor is added to a model, and decrease more for an important one.

A Plot of Model Fit

If the model predicts that 10% of the owners who live 1 km from a safe well and have 100 mg/l of arsenic will switch, then we'd like 10% of the owners in the data with those conditions to switch.

then predicted fraction = observed fraction at **X**

Cut the fitted values p into G intervals.

Compute the fraction f_i of Y=1's in each interval.

Plot f_i against the mean p_i for the interval

confidence interval for the sample mean:

$$\bar{p}_i \pm z_{\alpha/2}\sqrt{\bar{p}_i\left(1-\bar{p}_i\right)/n_i}$$

Sometimes called a calibration plot.

Calibration Plot For Well Model

predicted fraction:

mean fitted value μ in each interval

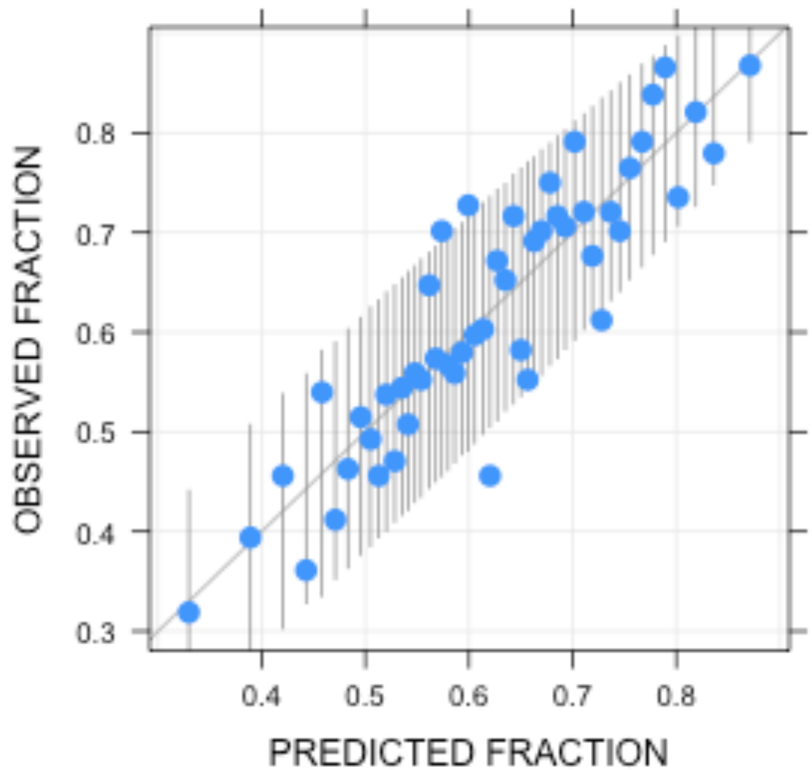
observed fraction:

mean Y in each interval

segments:

$$\bar{\mu}_i \pm z_{\alpha/2}\sqrt{\bar{\mu}_i(1-\bar{\mu}_i)/n_i}$$

n = #points in the interval



Segments show approximate 95% intervals. 50 intervals so expect ≈ 3 points outside their intervals.

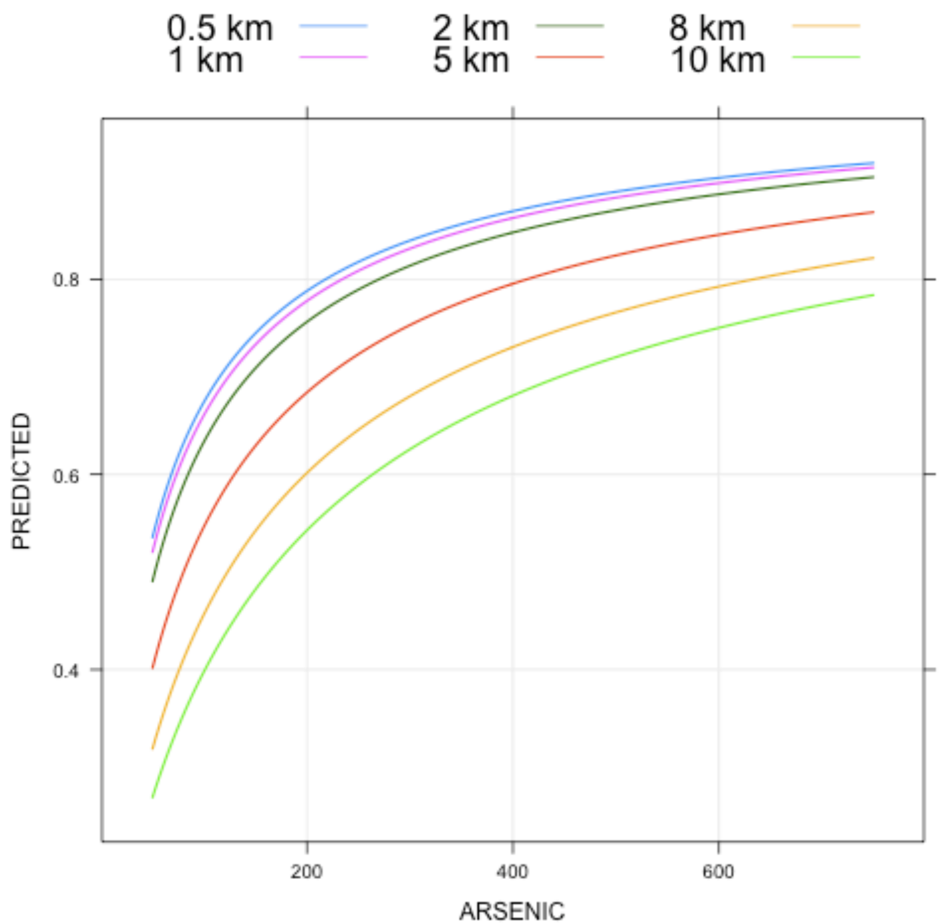
Plotting A Fitted Model

With no interaction, plot fitted vs X_1 for some values of X_2 (or vice versa)

- Use the original scale for arsenic for plotting, so the plot is easier to read.

  $b_{walk} = -.12$

  $b_{log(arsenic)} = .84$



lines represent different distances to a safe well

Uncertainty Around the Line

Repeat what we did for linear regression

- the coefficients are approx. multivariate normal
  - sample new coefficients
  - get new linear predictors
  - use `plogis` to translate to the probability scale