

BIOS 621 / 821 Session 2

Linear and Logistic Regression as Generalized Linear Models

Levi Waldron

CUNY SPH

**Learning
Objectives
and Outline**

Review of
multiple linear
regression

Linear
Regression as
a GLM

Logistic
Regression as
a GLM

Residuals for
logistic
regression

Likelihood
and
hypothesis
testing

Additive
vs. Multiplica-
tive
models

Learning Objectives and Outline

**Learning
Objectives
and Outline**

Review of
multiple linear
regression

Linear
Regression as
a GLM

Logistic
Regression as
a GLM

Residuals for
logistic
regression

Likelihood
and
hypothesis
testing

Additive
vs. Multiplica-
tive
models

Learning objectives

- define generalized linear models (GLM)
- define linear and logistic regression as special cases of GLMs
- distinguish between additive and multiplicative models
- define Pearson and deviance residuals
- describe application of the Wald test

**Learning
Objectives
and Outline**

Review of
multiple linear
regression

Linear
Regression as
a GLM

Logistic
Regression as
a GLM

Residuals for
logistic
regression

Likelihood
and
hypothesis
testing

Additive
vs. Multiplica-
tive
models

Outline

- Brief overview of multiple regression (Vittinghoff 4.1-4.3)
- Linear Regression as a Generalized Linear Model (Vittinghoff 4.1-4.3)
- Statistical inference for logistic regression (Vittinghoff 5.1-5.3)

Review of multiple linear regression

Systematic component

$$E[y|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

- x_p are the predictors or independent variables
- y is the outcome, response, or dependent variable
- $E[y|x]$ is the expected value of y given x
- β_p are the regression coefficients

Systematic plus random component

$$y_i = E[y|x] + \epsilon_i$$

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

Assumption: $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2)$

- Normal distribution
- Mean zero at every value of predictors
- Constant variance at every value of predictors
- Values that are statistically independent

Learning
Objectives
and Outline

Review of
multiple linear
regression

**Linear
Regression as
a GLM**

Logistic
Regression as
a GLM

Residuals for
logistic
regression

Likelihood
and
hypothesis
testing

Additive
vs. Multiplica-
tive
models

Linear Regression as a GLM

Generalized Linear Models (GLM)

- Linear regression is a special case of a broad family of models called “Generalized Linear Models” (GLM)
- This unifying approach allows to fit a large set of models using maximum likelihood estimation methods (MLE) (Nelder & Wedderburn, 1972)
- Can model many types of data directly using appropriate distributions, e.g. Poisson distribution for count data
- Transformations of Y not needed

Components of GLM

- **Random component** specifies the conditional distribution for the response variable
 - doesn't have to be normal
 - can be any distribution in the “exponential” family of distributions
- **Systematic component** specifies linear function of predictors (linear predictor)
- **Link** [denoted by $g(\cdot)$] specifies the relationship between the expected value of the random component and the systematic component
 - can be linear or nonlinear

Linear Regression as GLM

- **The model:**

$$y_i = E[y|x] + \epsilon_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \epsilon_i$$

- **Random component** of y_i is normally distributed:

$$\epsilon_i \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2)$$

- **Systematic component** (linear predictor):

$$\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}$$

- **Link function** here is the *identity link*:

$g(E(y|x)) = E(y|x)$. We are modeling the mean directly, no transformation.

Learning
Objectives
and Outline

Review of
multiple linear
regression

Linear
Regression as
a GLM

**Logistic
Regression as
a GLM**

Residuals for
logistic
regression

Likelihood
and
hypothesis
testing

Additive
vs. Multiplica-
tive
models

Logistic Regression as a GLM

The logistic regression model

- **The model:**

$$\text{Logit}(P(x)) = \log \left(\frac{P(x)}{1 - P(x)} \right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}$$

- **Random component:** y_i follows a Binomial distribution (outcome is a binary variable)
- **Systematic component:** linear predictor

$$\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}$$

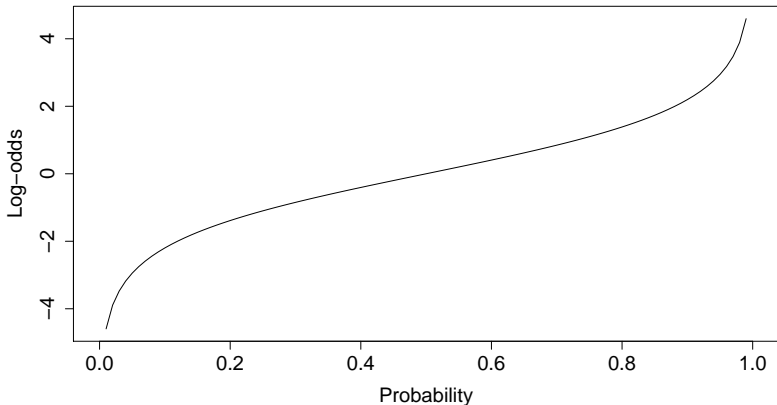
- **Link function:** *logit* (log of the odds that the event occurs)

$$g(P(x)) = \text{logit}(P(x)) = \log \left(\frac{P(x)}{1 - P(x)} \right)$$

$$P(x) = \sigma^{-1}(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi})$$

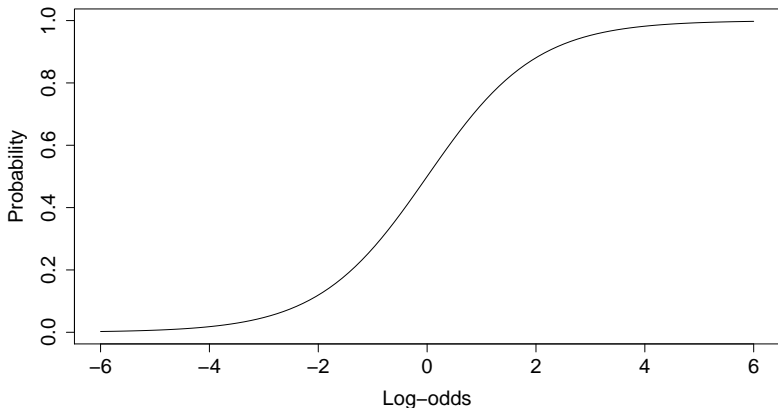
The logit function

```
logit <- function(P) log(P/(1-P))  
plot(logit, xlab="Probability", ylab="Log-odds",  
      cex.lab=1.5, cex.axis=1.5)
```



Inverse logit function

```
invLogit <- function(x) 1/(1+exp(-x))
```



Example: contraceptive use data

Load the contraceptive use data

```
suppressPackageStartupMessages(library(dplyr))
cuse <- read.table("cuse.dat", header=TRUE)
cuse <- mutate(cuse, percentusing = using / (using + notUsing) * 100) %>%
  mutate(n = using + notUsing)
cuse
```

age	education	wantsMore	notUsing	using	percentusing	n
<25	low	yes	53	6	10.16949	59
<25	low	no	10	4	28.57143	14
<25	high	yes	212	52	19.69697	264
<25	high	no	50	10	16.66667	60
25-29	low	yes	60	14	18.91892	74
25-29	low	no	19	10	34.48276	29
25-29	high	yes	155	54	25.83732	209
25-29	high	no	65	27	29.34783	92
30-39	low	yes	112	33	22.75862	145
30-39	low	no	77	80	50.95541	157
30-39	high	yes	118	46	28.04878	164
30-39	high	no	68	78	53.42466	146
40-49	low	yes	35	6	14.63415	41
40-49	low	no	46	48	51.06383	94
40-49	high	yes	8	8	50.00000	16
40-49	high	no	12	31	72.09302	43

Table One

Stratified by age				
		Overall	<25	25-29
##	n	16	4	4
##	education = low (%)	8 (50.0)	2 (50.0)	2 (50.0)
##	wantsMore = yes (%)	8 (50.0)	2 (50.0)	2 (50.0)
##	percentusing (mean (SD))	32.92 (17.51)	18.78 (7.64)	27.15 (6.53)
Stratified by age				
		30-39	40-49	
##	n	4	4	
##	education = low (%)	2 (50.0)	2 (50.0)	
##	wantsMore = yes (%)	2 (50.0)	2 (50.0)	
##	percentusing (mean (SD))	38.80 (15.65)	46.95 (23.82)	

See tableone vignette for e.g. how to export to Word / Excel

<https://cran.r-project.org/web/packages/tableone/vignettes/introduction.html>

Perform regression

```
fit1 <- glm(cbind(using, notUsing) ~ age + education + wantsMore,
            data=cuse, family=binomial("logit"))
summary(fit1)

##
## Call:
## glm(formula = cbind(using, notUsing) ~ age + education + wantsMore,
##      family = binomial("logit"), data = cuse)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.5148  -0.9376   0.2408   0.9822   1.7333
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.8082     0.1590  -5.083 3.71e-07 ***
## age25-29      0.3894     0.1759   2.214 0.02681 *
## age30-39      0.9086     0.1646   5.519 3.40e-08 ***
## age40-49      1.1892     0.2144   5.546 2.92e-08 ***
## educationlow  -0.3250     0.1240  -2.620 0.00879 **
## wantsMoreyes  -0.8330     0.1175  -7.091 1.33e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 165.772  on 15  degrees of freedom
## Residual deviance:  29.917  on 10  degrees of freedom
## AIC: 113.43
##
## Number of Fisher Scoring iterations: 4
```

Learning
Objectives
and Outline

Review of
multiple linear
regression

Linear
Regression as
a GLM

Logistic
Regression as
a GLM

**Residuals for
logistic
regression**

Likelihood
and
hypothesis
testing

Additive
vs. Multiplica-
tive
models

Residuals for logistic regression

Pearson residuals for logistic regression

- Traditional residuals $y_i - E[y_i|x_i]$ don't make sense for binary y .
- One alternative is *Pearson residuals*
 - take the difference between observed and fitted values (on probability scale 0-1), and divide by the standard deviation of the observed value.
- Let \hat{y}_i be the best-fit predicted probability for each data point, i.e. $g^{-1}(\beta_0 + \beta_1 x_{1i} + \dots)$
- y_i is the observed value, either 0 or 1.

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\text{Var}(\hat{y}_i)}}$$

Summing the squared Pearson residuals produces the *Pearson Chi-squared statistic*:

Deviance residuals for logistic regression

- Deviance residuals and Pearson residuals converge for high degrees of freedom
- Deviance residuals indicate the contribution of each point to the model *likelihood*
- Definition of deviance residuals:

$$d_i = s_i \sqrt{-2(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i))}$$

Where $s_i = 1$ if $y_i = 1$ and $s_i = -1$ if $y_i = 0$.

- Summing the deviances gives the overall deviance:
$$D = \sum_i d_i^2$$

Learning
Objectives
and Outline

Review of
multiple linear
regression

Linear
Regression as
a GLM

Logistic
Regression as
a GLM

Residuals for
logistic
regression

**Likelihood
and
hypothesis
testing**

Additive
vs. Multiplica-
tive
models

Likelihood and hypothesis testing

What is likelihood?

- The *likelihood* of a model is the probability of the observed outcomes given the model, sometimes written as:
 - $L(\theta|data) = P(data|\theta)$.
- Deviance residuals and the difference in log-likelihood between two models are related by:

$$\Delta(D) = -2 * \Delta(\log \text{ likelihood})$$

Likelihood Ratio Test

- Use to assess whether the reduction in deviance provided by a more complicated model indicates a better fit
- It is equivalent of the nested Analysis of Variance is a nested Analysis of Deviance
- The difference in deviance under H_0 is *chi-square distributed*, with df equal to the difference in df of the two models.

Likelihood Ratio Test (cont'd)

Learning
Objectives
and Outline

Review of
multiple linear
regression

Linear
Regression as
a GLM

Logistic
Regression as
a GLM

Residuals for
logistic
regression

**Likelihood
and
hypothesis
testing**

Additive
vs. Multiplica-
tive
models

```
fit0 <- glm(cbind(using, notUsing) ~ -1, data=cuse,  
            family=binomial("logit"))  
anova(fit0, fit1, test="LRT")
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
16	389.85387	NA	NA	NA
10	29.91722	6	359.9366	0

Wald test for individual regression coefficients

- Can use partial Wald test for a single coefficient:
 - $\frac{\hat{\beta}}{\sqrt{\text{var}(\hat{\beta})}} \sim t_{n-1}$
 - $\frac{(\hat{\beta} - \beta_0)^2}{\text{var}(\hat{\beta})} \sim \chi^2_{df=1}$ (large sample)
- Wald CI for β : $\hat{\beta} \pm t_{1-\alpha/2, n-1} \sqrt{\text{var}(\hat{\beta})}$
- Wald CI for odds-ratio: $e^{\hat{\beta} \pm t_{1-\alpha/2, n-1} \sqrt{\text{var}(\hat{\beta})}}$

Note: Wald test confidence intervals on coefficients can provide poor coverage in some cases, even with relatively large samples

Additive vs. Multiplicative models

Additive vs. Multiplicative models

- Linear regression is an *additive* model
 - e.g. for two binary variables $\beta_1 = 1.5$, $\beta_2 = 1.5$.
 - If $x_1 = 1$ and $x_2 = 1$, this adds 3.0 to $E(y|x)$
- Logistic regression is a *multiplicative* model
 - If $x_1 = 1$ and $x_2 = 1$, this adds 3.0 to $\log(\frac{P}{1-P})$
 - Odds-ratio $\frac{P}{1-P}$ increases 20-fold: $\exp(1.5 + 1.5)$ or $\exp(1.5) * \exp(1.5)$