Lecture 29: Logistic Regression

Pratheepa Jeganathan

12/04/2019

Recap

- What is a regression model?
- Descriptive statistics – graphical
- Descriptive statistics – numerical
- Inference about a population mean
- Difference between two population means
- Some tips on R
- Simple linear regression (covariance, correlation, estimation, geometry of least squares)
  - Inference on simple linear regression model
  - Goodness of fit of regression: analysis of variance.
  - F-statistics.
  - Residuals.
  - Diagnostic plots for simple linear regression (graphical methods).

Recap

- Multiple linear regression
  - Specifying the model.
  - Fitting the model: least squares.
  - Interpretation of the coefficients.
  - Matrix formulation of multiple linear regression
  - Inference for multiple linear regression
    - T-statistics revisited.
    - More F statistics.
    - Tests involving more than one β.
- Diagnostics – more on graphical methods and numerical methods
  - Different types of residuals
  - Influence
  - Outlier detection
  - Multiple comparison (Bonferroni correction)
  - Residual plots:
    - partial regression (added variable) plot,
    - partial residual (residual plus component) plot.

Recap

- Adding qualitative predictors
  - Qualitative variables as predictors to the regression model.
  - Adding interactions to the linear regression model.
  - Testing for equality of regression relationship in various subsets of a population
- ANOVA
  - All qualitative predictors.
  - One-way layout
  - Two-way layout
- Transformation
  - Achieving linearity
  - Stabilize variance
  - Weighted least squares
- Correlated Errors
  - Generalized least squares
- Bootstrapping linear regression
- Selection

Recap

- Colliniarity
  - Bias-variance tradeoff
  - Penalized Regression
    - Ridge
    - LASSO
    - Elastic net

Outline (Logistic regression)

- Most models so far have had response Y as continuous.
- Binary outcomes
  - Many responses in practice fall into the YES/NO framework.
  - Examples:
    1. medical: presence or absence of cancer
    2. financial: bankrupt or solvent
    3. industrial: passes a quality control test or not

Modelling probabilities

- For $0-1$ responses we need to model

$$\pi(x_1, \ldots, x_p) = P(Y = 1 | X_1 = x_1, \ldots, X_p = x_p)$$

- That is, Y is Bernoulli with a probability that depends on covariates $\{X_1, \ldots, X_p\}$.
- **Note:** $\text{Var}(Y) = \pi(1-\pi) = E(Y) \cdot (1 - E(Y))$
- **Or,** the binary nature forces a relation between mean and variance of Y.
- This makes logistic regression a `Generalized Linear Model`.

Flu shot example

- A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications, to get a flu shot in time for protection against an expected flu epidemic.
- In a pilot follow-up study, 50 clients were randomly selected and asked whether they actually received a flu shot. $Y = \texttt{Shot}$
- In addition, data were collected on their age $X_1 = \texttt{Age}$ and their health awareness $X_2 = \texttt{Health.Aware}$

A possible model

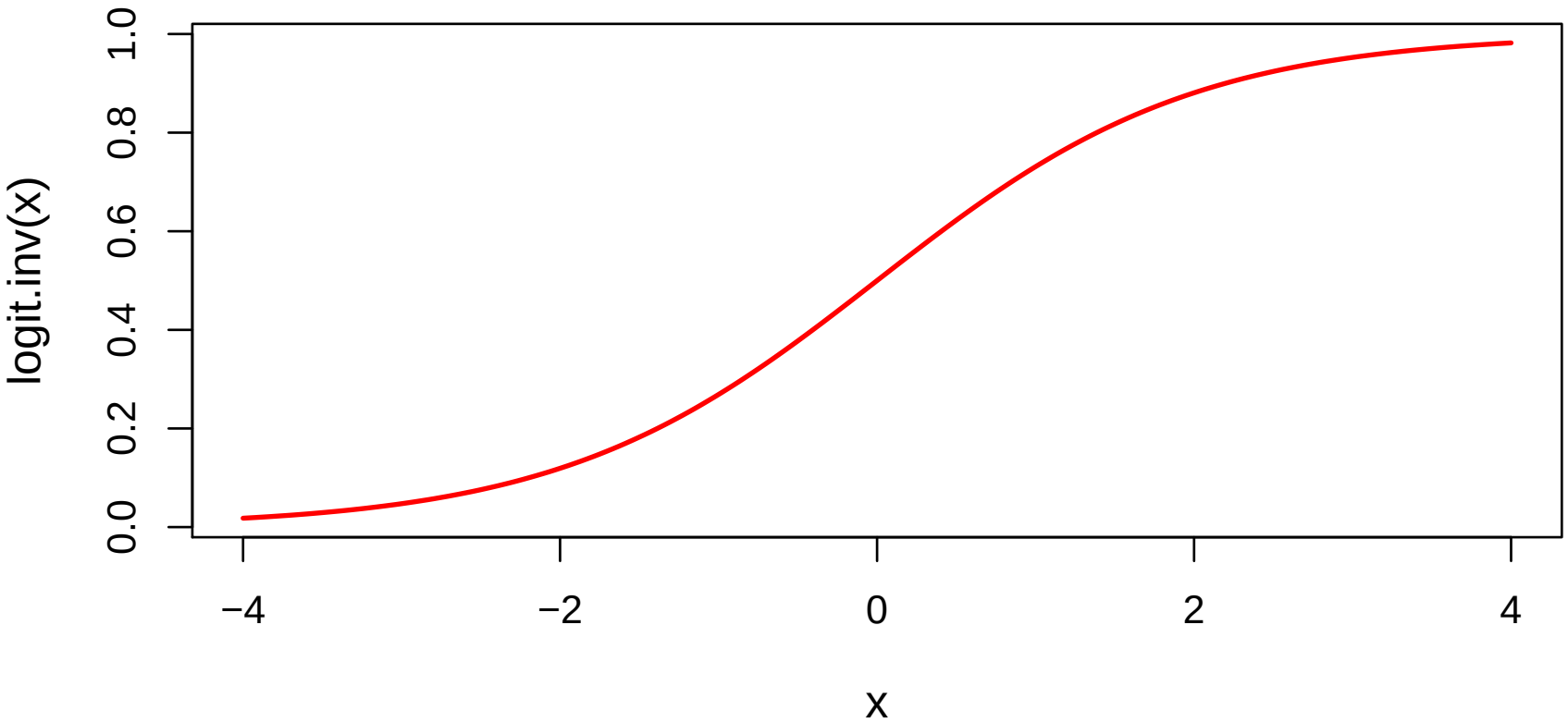
- Simplest model $\pi(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- Problems / issues:
  - We must have $0 \leq E(Y) = \pi(X_1, X_2) \leq 1$. OLS will not force this.
  - Ordinary least squares will not work because of relation between mean and variance.

Logistic model

- Logistic model $\pi(X_1, X_2) = \frac{\exp(\beta_0+\beta_1 X_1+\beta_2 X_2)}{1+\exp(\beta_0+\beta_1 X_1+\beta_2 X_2)}$
- This automatically fixes $0 \leq E(Y) = \pi(X_1, X_2) \leq 1$.
- **Define:**
  $\text{logit}(\pi(X_1, X_2)) = \log\left(\frac{\pi(X_1,X_2)}{1-\pi(X_1,X_2)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

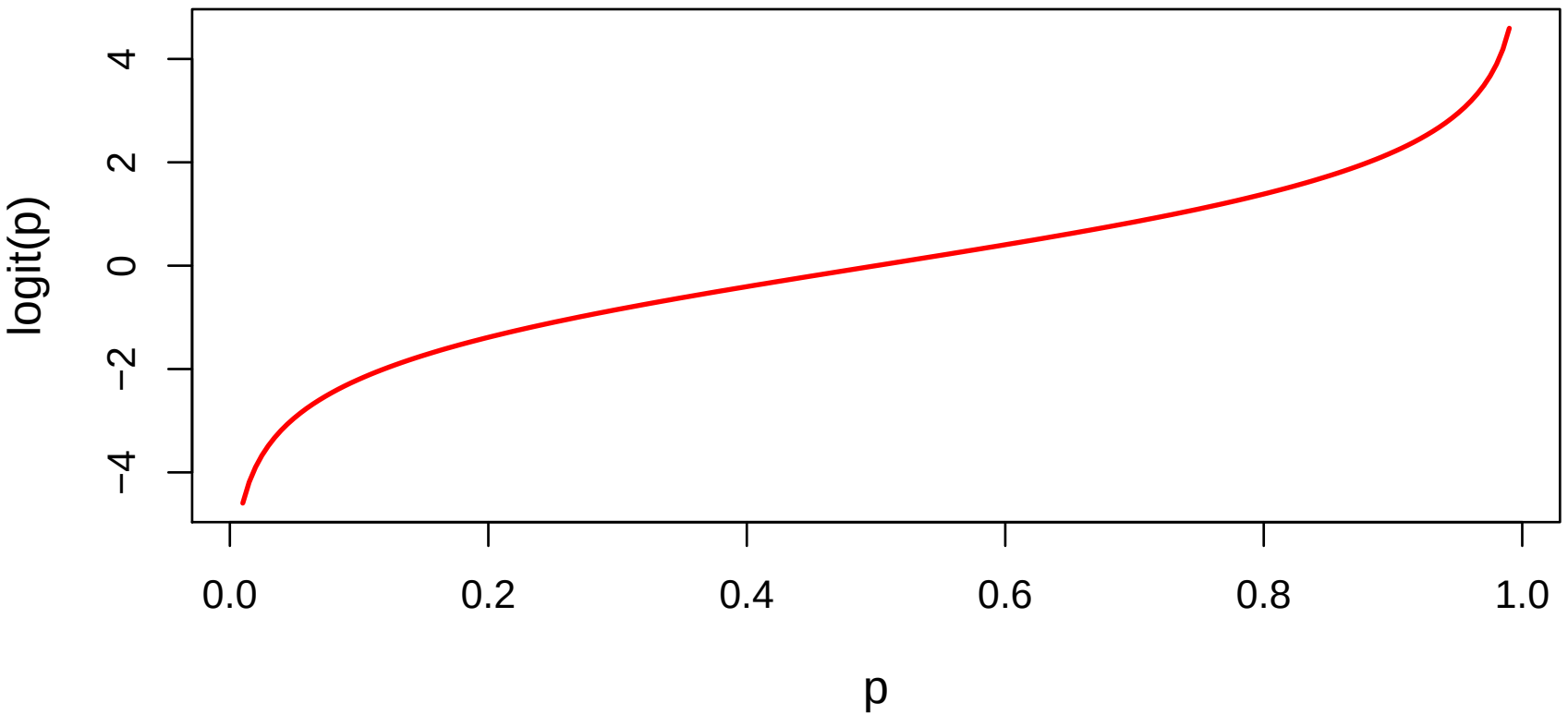
Logistic distribution

```
logit.inv = function(x) {
  return(exp(x) / (1 + exp(x)))
}
x = seq(-4, 4, length=200)
plot(x, logit.inv(x), lwd=2, type='l',
     col='red', cex.lab=1.2)
```



Logistic transform: logit

```r
logit = function(p) {
  return(log(p / (1 - p)))
}
p = seq(0.01, 0.99,length=200)
plot(p, logit(p), lwd=2, type='l',
     col='red', cex.lab=1.2)
```



Binary regression models

- Models $E(Y)$ as $F(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$ for some increasing function F (usually a distribution function).
- The logistic model uses the function (we called `logit.inv` above)

$$F(x) = \frac{e^x}{1 + e^x}.$$

- Can be fit using Maximum Likelihood / Iteratively Reweighted Least Squares.
- For logistic regression, coefficients have nice interpretation in terms of `odds ratios` (to be defined shortly).
- What about inference?

Criterion used to fit model

- Instead of sum of squares, logistic regression uses *deviance*:
- Let L be the likelihood function.
- $DEV(\mu|Y) = -2 \log L(\mu|Y) + 2 \log L(Y|Y)$ where μ is a location estimator for Y.
  - $L(Y|Y) = 1$
  - deviance is always larger or equal than zero
  - deviance is zero only if the fit is perfect
- If Y is Gaussian with independent $N(\mu_i, \sigma^2)$ entries then $DEV(\mu|Y) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \mu_i)^2$
- If Y is a binary vector, with mean vector π then $DEV(\pi|Y) = -2 \sum_{i=1}^{n} (Y_i \log(\pi_i) + (1 - Y_i) \log(1 - \pi_i))$

Minimizing deviance $\iff$ Maximum Likelihood

Deviance for logistic regression

- For any binary regression model, $\pi = \pi(\beta)$.
- The deviance is:

$$DEV(\beta|Y) = -2\sum_{i=1}^{n}(Y_i \text{logit}(\pi_i(\beta)) + \log(1-\pi_i(\beta)))$$

- For the logistic model, the RHS is:

$$-2\left[(X\beta)^T y + \sum_{i=1}^{n}\log\left(1+\exp\left(\sum_{j=1}^{p}X_{ij}\beta_j\right)\right)\right]$$

- The logistic model is special in that $\text{logit}(\pi(\beta)) = X\beta$. If we used a different transformation, the first part would not be linear in $X\beta$.
- *For ease of notation, we assume that* `X[,1]=1` *corresponding to* β_0

Flu shot example

- Response: Flu shot is taken or not.
- Predictors: Age, Health awareness

```
flu.table = read.table('http://stats191.stanford.edu/data/f
                       header=TRUE)
head(flu.table)
```

```
##   Shot Age Health.Aware
## 1    0  38           40
## 2    1  52           60
## 3    0  41           36
## 4    1  46           59
## 5    1  41           70
## 6    0  43           49
```

```
flu.glm = glm(Shot ~ Age + Health.Aware,
              data=flu.table,
              family=binomial())
```

Flu shot example

```
Call:
glm(formula = Shot ~ Age + Health.Aware, family = binomial(),
    data = flu.table)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5522  -0.2962  -0.1124   0.4208   2.3244

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -21.58458    6.41824  -3.363 0.000771 ***
Age           0.22178    0.07436   2.983 0.002858 **
Health.Aware  0.20351    0.06273   3.244 0.001178 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.029  on 49  degrees of freedom
Residual deviance: 32.416  on 47  degrees of freedom
AIC: 38.416

Number of Fisher Scoring iterations: 6
```

- null deviance: how well the reponse is predicted when there is no predictors
- residual deviance: how well the reponse is predicted when there predictors
- in this example, residual deviance is smaller than the null deviance.

Odds Ratios

- One reason logistic models are popular is that the parameters have simple interpretations in terms of **odds**

$$ODDS(A) = \frac{P(A)}{1 - P(A)}.$$

- Logistic model:

$$OR_{X_j} = \frac{ODDS(Y = 1|\ldots, X_j = x_j + h, \ldots)}{ODDS(Y = 1|\ldots, X_j = x_j, \ldots)} = e^{h\beta_j}$$

- If $X_j \in 0, 1$ is dichotomous, then odds for group with $X_j = 1$ are e^{β_j} higher, other parameters being equal.

Rare disease hypothesis

- When incidence is rare, $P(Y = 0) \approxeq 1$ no matter what the covariates X_j's are.
- In this case, odds ratios are almost ratios of probabilities:

$$OR_{X_j} \approxeq \frac{\mathbb{P}(Y = 1|\ldots, X_j = x_j + 1, \ldots)}{\mathbb{P}(Y = 1|\ldots, X_j = x_j, \ldots)}$$

- Hypothetical example: in a lung cancer study, if X_j is an indicator of smoking or not, a β_j of 5 means for smoking vs. non-smoking, smokers are $e^5 \approx 150$ times more likely to develop lung cancer

Flu shot example

- In flu example, the odds ratio for a 45 year old with health awareness 50 compared to a 35 year old with the same health awareness are

$$e^{-1.429284+3.647052} = 9.18$$

```
logodds = predict(flu.glm,
                  list(Age=c(35,45),Health.Aware=c(50,50)),
                  type='link')
logodds
```

```
##          1          2
## -3.647052 -1.429284
```

```
OR = exp(logodds[2])/exp(logodds[1]); OR
```

```
##        2
## 9.186801
```

Flu shot example

- The estimated probabilities are below, yielding a ratio of $0.1932/0.0254 \approx 7.61$. Not too far from 9.18.

```
prob = exp(logodds)/(1+exp(logodds))
prob
```

```
##         1         2
## 0.0254056 0.1932103
```

```
prob[2] / prob[1]
```

```
##        2
## 7.605027
```

Iteratively reweighted least squares

An algorithm to fit the model

1. Initialize $\widehat{\pi}_i = \bar{Y}, 1 \leq i \leq n$
2. Define
$$Z_i = g(\widehat{\pi}_i) + g'(\widehat{\pi}_i)(Y_i - \widehat{\pi}_i),$$
where g is the logit function.
3. Fit weighted least squares model
$$Z_i \sim \sum_{j=1}^{p} \beta_j X_{ij}, \qquad w_i = \widehat{\pi}_i(1 - \widehat{\pi}_i)$$
4. Set $\widehat{\pi}_i = \text{logit}^{-1}\left(\widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j X_{ij}\right)$.
5. Repeat steps 2-4 until convergence.

Newton-Raphson

- The Newton-Raphson updates for logistic regression are (minimizing Deviance)

$$\hat{\beta} \mapsto \hat{\beta} - \nabla^2 DEV(\hat{\beta})^{-1} \nabla DEV(\hat{\beta})$$

- These turn out to be the same as the updates above (as in the Iteratively reweighted least squares).
- In earlier statistical software one might only have access to a weighted least squares estimator.

Inference

- One thing the IRLS procedure hints at is what the approximate limiting distribution is.
  - The IRLS procedure suggests using approximation $\widehat{\beta} \approx N(\beta, (X^T W X)^{-1})$
  - This allows us to construct CIs, test linear hypotheses, etc.
  - Intervals formed this way are called *Wald intervals*.

Flu shot example

- 95% CI for β_1

```
center = coef(flu.glm)['Age']
SE = sqrt(vcov(flu.glm)['Age', 'Age'])
U = center + SE * qnorm(0.975)
L = center - SE * qnorm(0.975)
data.frame(L, center, U)
```

```
##                L    center         U
## Age 0.0760395 0.2217768 0.3675141
```

Covariance

- The estimated covariance `vcov(flu.glm)` uses the weights computed from the fitted model.

```
pi.hat = fitted(flu.glm)
W.hat = pi.hat * (1 - pi.hat)
X = model.matrix(flu.glm)
C = solve(t(X) %*% (W.hat * X))
c(SE, sqrt(C['Age', 'Age']))
```

```
## [1] 0.07435712 0.07435807
```

Confidence intervals in R

- The intervals above are slightly different from what R will give you if you ask it for confidence intervals.
- R uses so-called profile intervals.
- For large samples the two methods should agree quite closely.

Confidence intervals in R

```
CI = confint(flu.glm)
CI
```

```
##                    2.5 %      97.5 %
## (Intercept)  -38.0402235 -11.6669218
## Age            0.1004533   0.4046856
## Health.Aware   0.1026984   0.3595480
```

```
# profile intervals are not symmetric around the estimate.
mean(CI[2,])
```

```
## [1] 0.2525694
```

```
# we computed center of the interval as follows
data.frame(L, center, U)
```

```
##                L    center         U
## Age 0.0760395 0.2217768 0.3675141
```

Testing in logistic regression

What about comparing full and reduced model?

- For a model $\mathcal{M}$, $DEV(\mathcal{M})$ replaces $SSE(\mathcal{M})$.
- In least squares regression (with σ^2 known), we use

$$\frac{1}{\sigma^2}\left(SSE(\mathcal{M}_R) - SSE(\mathcal{M}_F)\right) \overset{H_0:\mathcal{M}_R}{\sim} \chi^2_{df_R - df_F}$$

- This is closely related to F with large df_F: approximately $F_{df_R - df_F, df_R} \cdot (df_R - df_F)$.
- For logistic regression this difference in SSE is replaced with

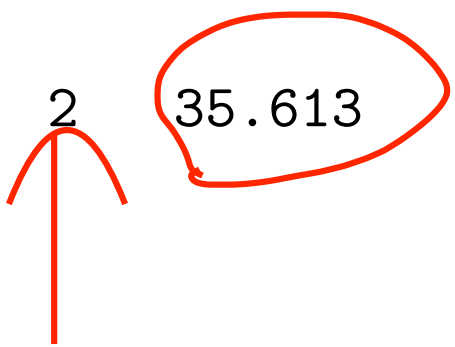
$$DEV(\mathcal{M}_R) - DEV(\mathcal{M}_F) \overset{n\to\infty, H_0:\mathcal{M}_R}{\sim} \chi^2_{df_R - df_F}$$

- Resulting tests do not agree numerically with those coming from IRLS (Wald tests). Both are often used.

Flu shot example

```
anova(glm(Shot ~ 1,
          data=flu.table,
          family=binomial()),
      flu.glm)
```

```
## Analysis of Deviance Table
##
## Model 1: Shot ~ 1
## Model 2: Shot ~ Age + Health.Aware
##   Resid. Df Resid. Dev Df Deviance
## 1        49     68.029
## 2        47     32.416  2   35.613
```



Flu shot example

```
anova(glm(Shot ~ Health.Aware,
          data=flu.table,
          family=binomial()),
      flu.glm)
```

```
## Analysis of Deviance Table
##
## Model 1: Shot ~ Health.Aware
## Model 2: Shot ~ Age + Health.Aware
##   Resid. Df Resid. Dev Df Deviance
## 1        48     49.279
## 2        47     32.416  1   16.863
```

Flu shot example

- We should compare this difference in deviance with a χ^2_1 random variable.

```
# testing ~1 vs ~1 + Health.Aware + Age
1 - pchisq(35.61, 2)
```

```
## [1] 1.850916e-08
```

```
# testing ~ 1 + Health.Aware vs ~1 + Health.Aware + Age
1 - pchisq(16.863, 1)
```

```
## [1] 4.017719e-05
```

- Let's compare this with the Wald test:

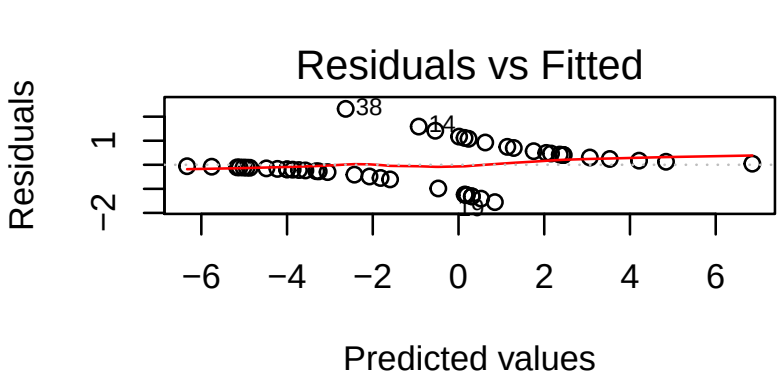
```
summary(flu.glm)
```

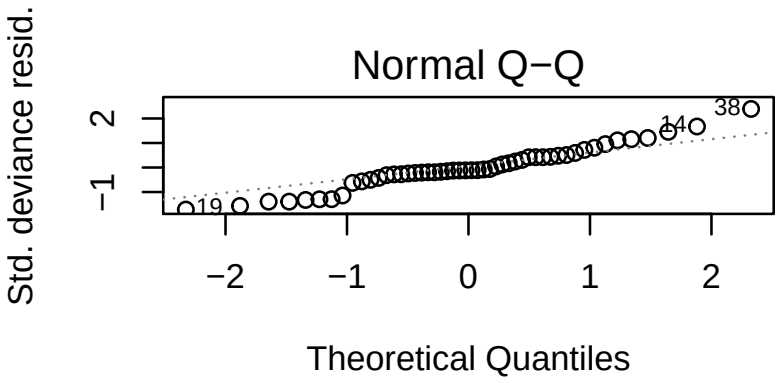
```
Call:
glm(formula = Shot ~ Age + Health.Aware, family = binomial(),
    data = flu.table)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5522  -0.2962  -0.1124   0.4208   2.3244

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -21.58458    6.41824  -3.363 0.000771 ***
Age           0.22178    0.07436   2.983 0.002858 **
Health.Aware  0.20351    0.06273   3.244 0.001178 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.029  on 49  degrees of freedom
Residual deviance: 32.416  on 47  degrees of freedom
AIC: 38.416

Number of Fisher Scoring iterations: 6
```

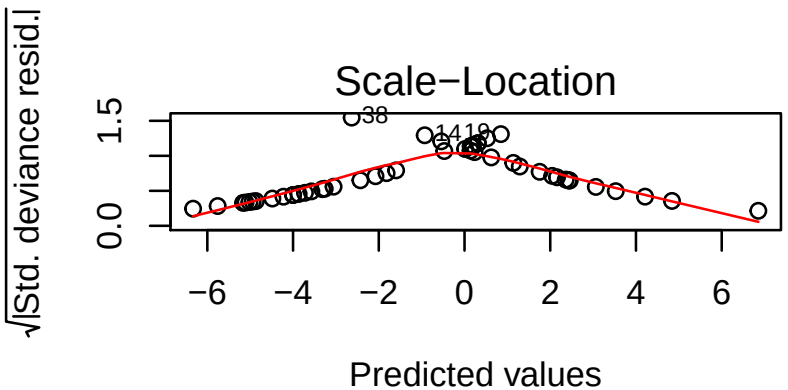
Diagnostics

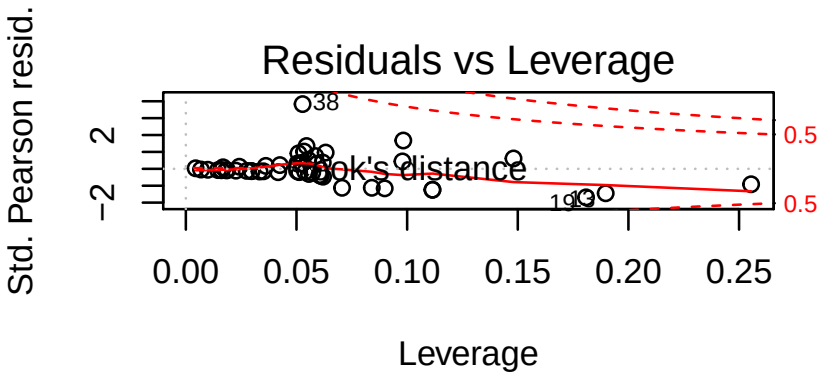
- Similar to least square regression, only residuals used are usually *deviance residuals* $r_i = \text{sign}(Y_i - \widehat{\pi}_i)\sqrt{DEV(\widehat{\pi}_i|Y_i)}$.
- These agree with usual residual for least square regression.

Diagnostics









Diagnostics

```
influence.measures(flu.glm)
```

	dfb.1_ <dbl>	dfb.Age <dbl>	dfb.Hl.A <dbl>	dffit <dbl>	cov.r <dbl>	cook.d <dbl>	hat <dbl>	inf <fctr>
1	-0.017208450	0.015425632	0.015642750	-0.017627295	1.0826918	3.659998e-05	0.016044175	
2	-0.102896560	0.095601518	0.104158961	0.134913568	1.1016580	2.242955e-03	0.052309317	
3	-0.015043331	0.012796775	0.014423117	-0.015557954	1.0812268	2.849979e-05	0.014540618	
4	-0.058889061	0.024180020	0.119204481	0.287549399	1.0400288	1.178645e-02	0.058501495	
5	-0.103487667	0.038531289	0.171669954	0.234752910	1.1440718	6.941304e-03	0.097887286	
6	-0.130919008	0.118975462	0.110976921	-0.152033795	1.1089962	2.859279e-03	0.060396517	
7	-0.066733843	0.066912148	0.060032141	0.077608468	1.1041718	7.216821e-04	0.042487429	
8	-0.043827442	0.043992127	0.034380158	-0.047042275	1.1010477	2.622724e-04	0.035080574	
9	-0.018883246	0.018809963	0.015150022	-0.019796169	1.0854748	4.617103e-05	0.018665882	
10	-0.112977727	0.096952606	0.132645407	0.208414096	1.0760659	5.683995e-03	0.055255614	
11	-0.009577574	0.095836356	-0.043048893	0.364897254	1.0037014	2.083004e-02	0.063175325	
12	0.020211669	-0.133730542	0.058121041	-0.435658880	0.9792425	3.195699e-02	0.070590655	
13	0.192769578	-0.507763090	0.117899074	-0.937781945	1.0402263	1.632619e-01	0.189724920	*
14	0.372025379	-0.472612878	-0.145485378	0.688135732	0.9022852	1.019381e-01	0.098269672	
15	-0.038715894	0.035962655	0.033505165	-0.040345463	1.0954845	1.927046e-04	0.029695080	
16	-0.057483528	0.064804695	0.037346215	-0.068923294	1.1213473	5.647229e-04	0.053933004	
17	-0.046411886	0.139137544	-0.108584803	-0.479730861	0.9930315	3.855770e-02	0.084036358	
18	-0.029026018	0.024985096	0.030041389	0.032568615	1.0900109	1.253857e-04	0.024164623	
19	0.111761811	0.199472599	-0.490738781	-1.006689601	0.9788611	2.098408e-01	0.180843476	*
20	-0.063892521	0.051985794	0.062463199	-0.069722982	1.1049785	5.804322e-04	0.041691779	
21	-0.148883584	0.116208572	0.142755961	-0.193270807	1.0962350	4.759595e-03	0.061993468	
22	-0.046450729	0.036647120	0.052203167	0.056584320	1.1005070	3.808388e-04	0.036145888	
23	-0.058937986	0.036950964	0.069332741	-0.075873231	1.1164261	6.867114e-04	0.051205093	
24	-0.046667990	0.040386222	0.043379257	-0.049176718	1.0983427	2.870168e-04	0.033293002	
25	0.181347490	-0.108435013	-0.178659012	0.430487951	0.9213015	3.496902e-02	0.054530052	
26	-0.009162371	0.007591158	0.009034704	-0.009529635	1.0767247	1.068087e-05	0.009944897	

1–26 of 50 rows

Previous 1 2 Next

Model selection

- As the model is a likelihood based model, each fitted model has an AIC.
- Stepwise selection can be used easily

Model selection

```
step(flu.glm, scope=list(upper= ~.^2),
     direction='both')
```

```
Start:  AIC=38.42
Shot ~ Age + Health.Aware

                   Df Deviance    AIC
+ Age:Health.Aware  1   24.283 32.283
<none>                  32.416 38.416
- Age               1   49.279 53.279
- Health.Aware      1   56.078 60.078

Step:  AIC=32.28
Shot ~ Age + Health.Aware + Age:Health.Aware

                   Df Deviance    AIC
<none>                  24.283 32.283
- Age:Health.Aware  1   32.416 38.416

Call:  glm(formula = Shot ~ Age + Health.Aware + Age:Health.Aware, family = binomial(),
    data = flu.table)

Coefficients:
     (Intercept)              Age      Health.Aware
        26.75944          -0.88151           -0.82239
Age:Health.Aware
         0.02365

Degrees of Freedom: 49 Total (i.e. Null);  46 Residual
Null Deviance:      68.03
Residual Deviance: 24.28        AIC: 32.28
```

Penalized regression

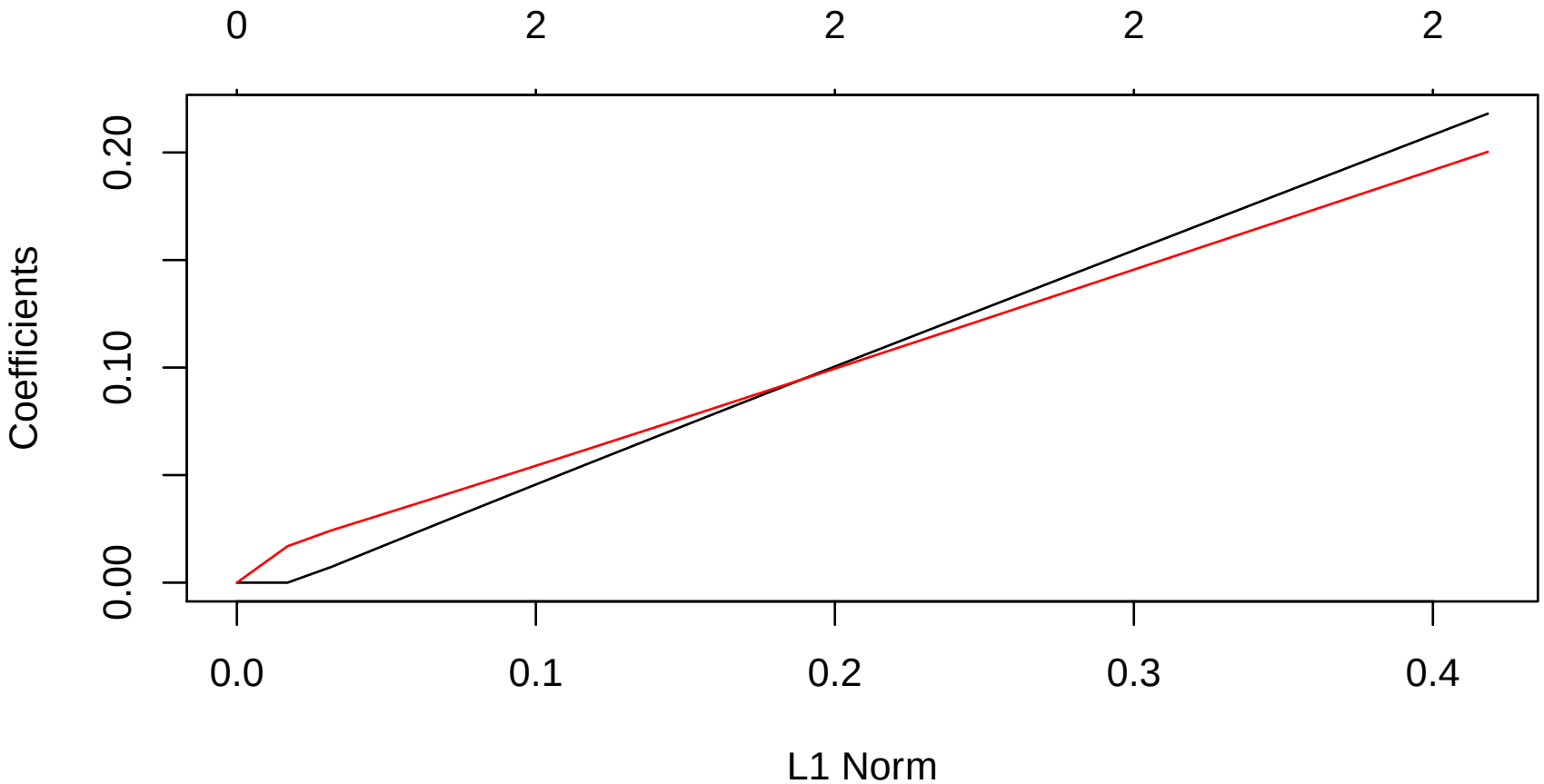
LASSO

- Instead of just minimizing deviance, we can also look at penalized versions

$$\text{minimize}_\beta \frac{1}{2n} DEV(\beta) + \lambda \|\beta\|_1$$

Flu shot example

```
library(glmnet)
X = model.matrix(flu.glm)[,-1]
Y = as.numeric(flu.table$Shot)
G = glmnet(X, Y, family="binomial")
plot(G)
```



Spam data set

- A data frame with 4601 observations on the 57 numeric predictor variables and a response `spam` is factor variable with 2 levels `email`, `spam`.
- More information of the data is here.

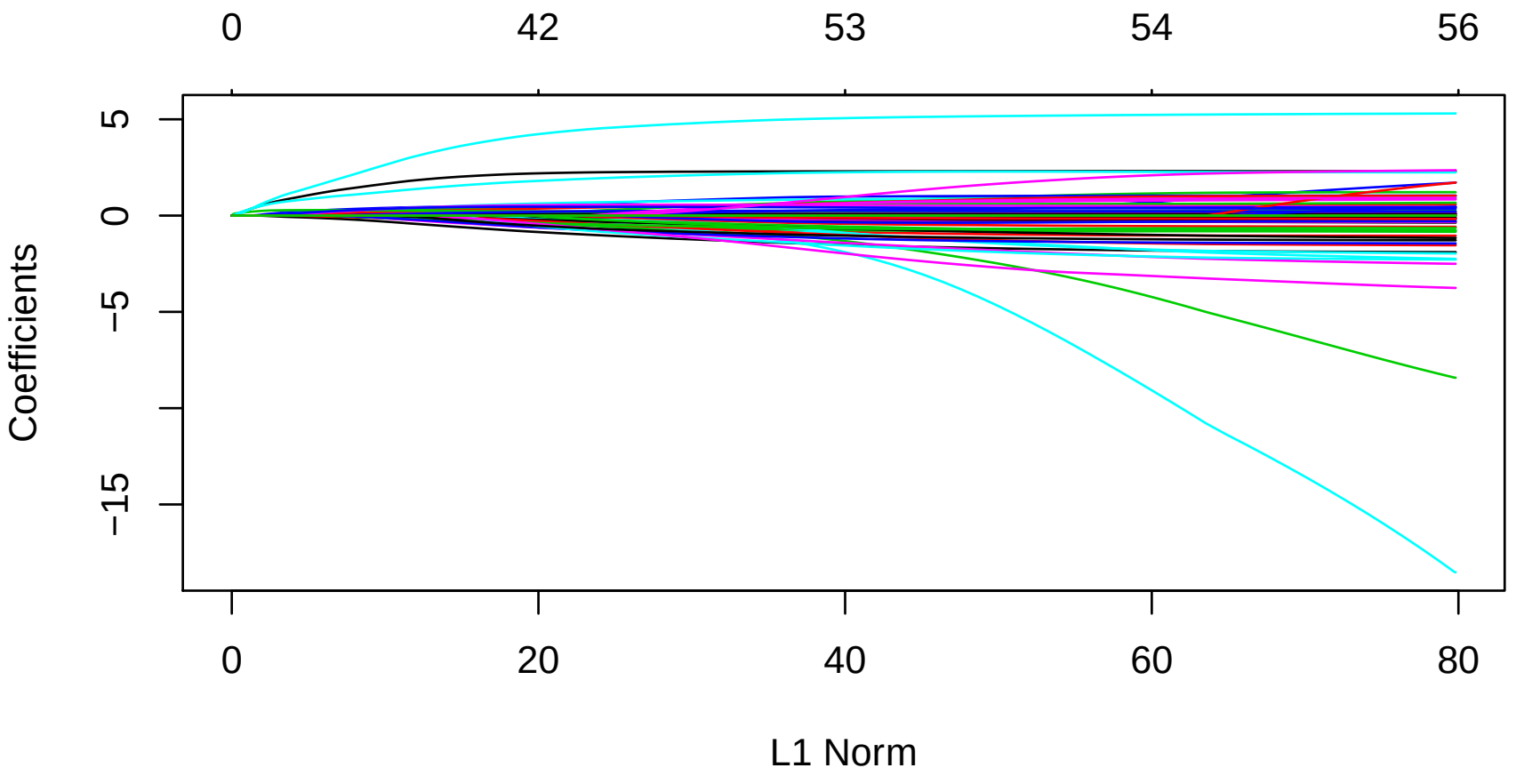
```
library(ElemStatLearn)
data(spam)
dim(spam)
```

```
## [1] 4601   58
```

```
X = model.matrix(spam ~ ., data=spam)[,-1]
Y = as.numeric(spam$spam == 'spam')
```

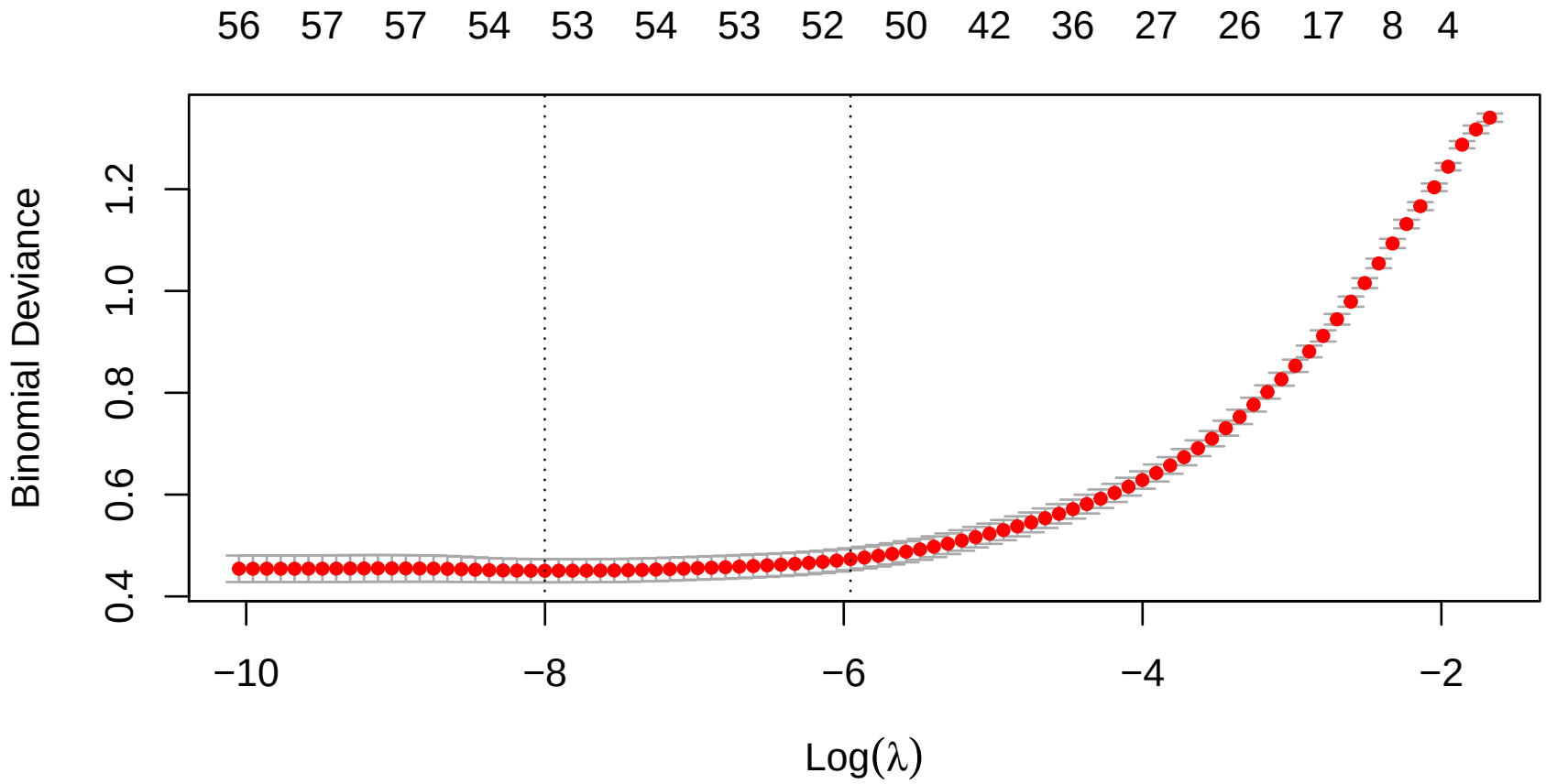
Spam data set

```
G = glmnet(X, Y, family='binomial')
plot(G)
```



Spam data set

```
CV = cv.glmnet(X, Y, family='binomial')
plot(CV)
```




```
c(CV$lambda.min, CV$lambda.1se)
```

```
## [1] 0.0003349517 0.0025934091
```

Extracting coefficients from glmnet

```
beta.hat = coef(G, s=CV$lambda.1se)
beta.hat
```

```
## 58 x 1 sparse Matrix of class "dgCMatrix"
##                              1
## (Intercept) -1.5752618121
## A.1         -0.1716645042
## A.2         -0.0871840559
## A.3          0.1318482487
## A.4          0.0936079556
## A.5          0.5081162007
## A.6          0.5482723371
## A.7          2.2785920925
## A.8          0.5519184177
## A.9          0.5002890226
## A.10         0.0829896113
## A.11         .
## A.12        -0.1157514617
```

Probit model

Probit model

- Probit regression model:

$$\Phi^{-1}(E(Y|X)) = \sum_{j=1}^{p} \beta_j X_j$$

where Φ is CDF of $N(0,1)$, i.e. $\Phi(t) = \texttt{pnorm(t)}$, $\Phi^{-1}(q) = \texttt{qnorm}(q)$.

- Regression function

$$\begin{aligned} E(Y|X) &= E(Y|X_1, \ldots, X_p) \\ &= P(Y = 1|X_1, \ldots, X_p) \\ &= \texttt{pnorm}\left(\sum_{j=1}^{p} \beta_j X_j\right) \end{aligned}$$

- In logit, probit and cloglog $\text{Var}(Y_i) = \pi_i(1 - \pi_i)$ but the model for the mean is different.

- Coefficients no longer have an odds ratio interpretation.

Probit model

```
summary(glm(Shot ~ Age + Health.Aware,
            data=flu.table,
            family=binomial(link='probit')))
```

```
Call:
glm(formula = Shot ~ Age + Health.Aware, family = binomial(link = "probit"),
    data = flu.table)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5471  -0.2883  -0.0648   0.4060   2.2955

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.35039    3.22797  -3.826 0.000130 ***
Age           0.12786    0.03887   3.289 0.001005 **
Health.Aware  0.11642    0.03237   3.596 0.000323 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.029  on 49  degrees of freedom
Residual deviance: 32.076  on 47  degrees of freedom
AIC: 38.076

Number of Fisher Scoring iterations: 7
```

Generalized linear models

Given a dataset $(Y_i, X_{i1}, \ldots, X_{ip}), 1 \leq i \leq n$ we consider a model for the distribution of $Y|X_1, \ldots, X_p$.

- If $\eta_i = g(E(Y_i|X_i)) = g(\mu_i) = \sum_{j=1}^{p} \beta_j X_{ij}$ then g is called the *link* function for the model.
- If $\text{Var}(Y_i) = \phi \cdot V(\mathbb{E}(Y_i)) = \phi \cdot V(\mu_i)$ for $\phi > 0$ and some function V, then V is the called *variance* function for the model.
- Canonical reference Generalized linear models.

Binary regression as GLM

- For a logistic model, $g(\mu) = \text{logit}(\mu), \qquad V(\mu) = \mu(1-\mu)$.
- For a probit model, $g(\mu) = \Phi^{-1}(\mu), \qquad V(\mu) = \mu(1-\mu)$.
- For a cloglog model,
  $g(\mu) = -\log(-\log(\mu)), \qquad V(\mu) = \mu(1-\mu)$.
- All of these have *dispersion* $\phi = 1$.

Reference

- **CH** Chapter 12.
- Lecture notes of Jonathan Taylor .