Outline

Logistic Regression

- Recall that the logistic regression specifies the model for a binary response variable Y_i as

$$Y_i \sim Ber(\pi_i), \quad \text{independently.}$$

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ is the $p \times 1$ vector of logistic regression coefficients.
- $\boldsymbol{X}_i = (1, X_{i,1}, \ldots, X_{i,p-1})'$ is the $p \times 1$ vector of explanatory variables of the ith observation.
- The mean model for logistic regression is

$$\mathbb{E}(Y_i) = \pi_i = \frac{\exp(\boldsymbol{X}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}_i'\boldsymbol{\beta})}.$$

- We use maximum likelihood for parameter estimation.

Likelihood Function

- Since $Y_i \sim Ber(\pi_i)$, the probability density function is

$$f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i},$$

  where $Y_i = 0$ or 1, $i = 1, \ldots, n$.

- Since Y_i's are independent, the joint probability density function is

$$f(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}.$$

Likelihood Function

- Take logarithm of $f(Y_1, \ldots, Y_n)$ and obtain

$$
\begin{aligned}
l(\boldsymbol{\beta}) &= \log f(Y_1, \ldots, Y_n) \\
&= \sum_{i=1}^{n} \{ Y_i \log(\pi_i) + (1 - Y_i) \log(1 - \pi_i) \} \\
&= \sum_{i=1}^{n} \left\{ Y_i \log\left(\frac{\pi_i}{1 - \pi_i} \right) + \log(1 - \pi_i) \right\} \\
&= \sum_{i=1}^{n} \left[Y_i(\boldsymbol{X}_i' \beta) - \log\{1 + \exp(\boldsymbol{X}_i' \beta)\} \right].
\end{aligned}
$$

- Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1})'$ denote the MLE of $\boldsymbol{\beta}$.

Fitting a binary regression GLM: IRLS

- Algorithm:
  1. Initialize: set $\hat{\mu}_i = 0.999$ or 0.001 depending on whether $Y_i = 1$ or 0.
  2. Compute $Z_i \to g(\hat{\mu}_i) + g'(\hat{\mu}_i)(Y_i - \hat{\mu}_i)$.
  3. Obtain $\hat{\beta}$ by regressing $\boldsymbol{Z}$ onto $\boldsymbol{X}$ using WLS with weights $W_i^{-1} = g'(\hat{\mu}_i)^2 V(\hat{\mu}_i)$ to
  4. Compute $\hat{\mu}_i = g^{-1}(\boldsymbol{X}_i' \hat{\beta})$.
  5. Repeat steps 2–4 until convergence.
- If ϕ has to be estimated, a simple choice is Pearson's X^2:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

- Approximate distribution of $\hat{\beta}$:

$$\hat{\beta} \sim N(\beta, \phi(\boldsymbol{X}^T \hat{W} \boldsymbol{X})^{-1}).$$

Large-Sample (Asymptotic) Properties of MLEs

- Inference about the logistic regression coefficients relies on asymptotic normality of the MLEs.
- Let β^0 denote the $p \times 1$ vector of true regression parameters.
- Let $\boldsymbol{H}$ denote the $p \times p$ **Hessian matrix** $\boldsymbol{H}(\beta) = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$.
- Let $\mathcal{I}(\beta^0)$ denote the $p \times p$ **Fisher information matrix** $\mathcal{I}(\beta) = -\mathbb{E}(\boldsymbol{H}(\beta))$ evaluated at β^0.

Approximate distribution of $\hat{\beta}$

Under suitable regularity conditions, as $n \to \infty$,

$$\hat{\beta} \approx N\left(\beta^0, \boldsymbol{V}\beta)\right), \quad \text{where} \quad \boldsymbol{V}(\hat{\beta}) = -\boldsymbol{H}(\hat{\beta})^{-1}.$$

Outline

Wald Test

- For individual regression coefficients,

$$\frac{\hat{\beta}_k - \beta_k}{S\{\hat{\beta}_k\}} \approx N(0,1), \quad k = 0, 1, \ldots, p-1,$$

  where $S^2\{\hat{\beta}_k\}$ is the kth diagonal element of the matrix $\boldsymbol{V}\{\hat{\boldsymbol{\beta}}\}$.
- To test $H_0 : \beta_k = 0$ versus $H_A : \beta_k \neq 0$, compute the observed statistic

$$z^* = \frac{\hat{\beta}_k}{s\{\hat{\beta}_k\}}$$

  and the decision rule is to reject H_0 if $|z^*| > z_{1-\alpha/2}$.
- The test above is also known as a **Wald test.**
- An approximate $1-\alpha$ confidence interval for β_k is

$$\hat{\beta}_k \pm z_{1-\alpha/2} s\{\hat{\beta}_k\}.$$

Outline

Likelihood Ratio Test

- The hypothesis of interest is

$$H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0.$$

- Let the full model have the logistic response function

$$\pi = E(Y) = \frac{\exp(\boldsymbol{X}_F'\boldsymbol{\beta}_F)}{1 + \exp(\boldsymbol{X}_F'\boldsymbol{\beta}_F)}$$

  where $\boldsymbol{X}_F = (1, X_1, \ldots, X_{p-1})'$ and $\boldsymbol{\beta}_F = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$.
- Under the H_0, let the reduced model have the logistic response function

$$\pi = E(Y) = \frac{\exp(\boldsymbol{X}_R'\boldsymbol{\beta}_R)}{1 + \exp(\boldsymbol{X}_R'\boldsymbol{\beta}_R)}$$

  where $\boldsymbol{X}_R = (1, X_1, \ldots, X_{q-1})'$ and
  $\boldsymbol{\beta}_R = (\beta_0, \beta_1, \ldots, \beta_{q-1})'$.

Likelihood Ratio Test

- Let $\mathcal{L}(F)$ denote the likelihood function evaluated at the MLE $\hat{\beta}_F$ under the full model.
- Let $\mathcal{L}(R)$ denote the likelihood function evaluated at the MLE $\hat{\beta}_R$ under the reduced model.
- The **likelihood ratio test (LRT)** statistic is defined as

$$\begin{aligned} G^2 &= -2\log\left\{\frac{\mathcal{L}(R)}{\mathcal{L}(F)}\right\} \\ &= -2\left\{\log\mathcal{L}(R) - \log\mathcal{L}(F)\right\} \overset{H_0}{\sim} \chi^2_{df_R - df_F} \end{aligned}$$

  where $df_R = n - q$, $df_F = n - p$, and thus $df_R - df_F = p - q$.
- The decision rule is to reject H_0 if $G^2 > \chi^2_{p-q,1-\alpha}$.

Example: disease outbreak

```
> mydata = read.table("disease.txt", header=T); attach(mydata)
> head(mydata)
  Case X1 X2 X3 X4 Y
1    1 33  0  0  0 0
2    2 35  0  0  0 0
3    3  6  0  0  0 0
4    4 60  0  0  0 0
5    5 18  0  1  0 1
6    6 26  0  1  0 0
> table(Y)
Y
0  1
67 31
> table(Y)/length(Y)
Y
0         1
0.6836735 0.3163265
```

Example: disease outbreak

A first-order multiple logistic regression model with the three predictor variables was considered *a priori* to be reasonable:
$\pi_i = (1 + \exp(-\boldsymbol{X}_i'\beta))^{-1}$.

```
> glm4 = glm(Y~., data=mydata[,-1], family=binomial("logit"))
> summary(glm4)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.31293    0.64259  -3.599 0.000319 ***
X1           0.02975    0.01350   2.203 0.027577 *
X2           0.40879    0.59900   0.682 0.494954
X3          -0.30525    0.60413  -0.505 0.613362
X4           1.57475    0.50162   3.139 0.001693 **
---
Null deviance: 122.32  on 97  degrees of freedom
Residual deviance: 101.05  on 93  degrees of freedom
AIC: 111.05
```

The fitted logistic response function is

$$\begin{aligned}\hat{\pi}_i &= \frac{1}{1 + \exp(-\boldsymbol{X}_i'\hat{\beta})} \\ &= \frac{1}{1 + \exp(2.313 - 0.02975X_{i,1} - 0.4088X_{i,2} + 0.3053X_{i,3} - 1.575X_{i,4})}.\end{aligned}$$

Example: disease outbreak

```
> ci95 = confint.default(glm4)
> round(cbind(summary(glm4)$coeff, ci95),3)
            Estimate Std. Error z value Pr(>|z|)  2.5 % 97.5 %
(Intercept)   -2.313      0.643  -3.599    0.000 -3.572 -1.053
X1             0.030      0.014   2.203    0.028  0.003  0.056
X2             0.409      0.599   0.682    0.495 -0.765  1.583
X3            -0.305      0.604  -0.505    0.613 -1.489  0.879
X4             1.575      0.502   3.139    0.002  0.592  2.558
> round(cbind(exp(glm4$coef), exp(ci95)),3)   ## odds ratios and CI
                  2.5 % 97.5 %
(Intercept) 0.099 0.028  0.349
X1          1.030 1.003  1.058
X2          1.505 0.465  4.869
X3          0.737 0.226  2.408
X4          4.830 1.807 12.909
```

- $\exp(\hat{\beta}_k)$: estimated odds ratio for X_k.
- $\hat{\beta}_1 = 0.030$ and $\exp \hat{\beta}_1 = 1.030$. The odds of a person having contracted the disease increase by about 3.0 percent with each additional year of age (X_1), for given socioeconomic status and city sector location.
- A 95% CI for $\exp \beta_1$ is (1.003, 1.058).

Example: disease outbreak

- Conduct LRTs to see whether a variable could be dropped from the logistic regression model.
- $H_0 : \beta_1 = 0$. The P-value of this test is .023. We conclude that X_1 should not be dropped from the model.

```
> anova(glm(Y~X2+X3+X4,  family=binomial("logit")), glm4, test="Chisq")
Analysis of Deviance Table

Model 1: Y ~ X2 + X3 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        94     106.20
2        93     101.05  1   5.1495  0.02325 *
> glm4$deviance
[1] 101.0542
> (glm(Y~X2+X3+X4,  family=binomial("logit")))$deviance
[1] 106.2037
```

106.2037 - 101.0542 = 5.1495

```
> library(car); Anova(glm4, type="III")
Analysis of Deviance Table (Type III tests)

Response: Y
   LR Chisq Df Pr(>Chisq)
X1   5.1495  1   0.023253 *
X2   0.4669  1   0.494416
X3   0.2560  1   0.612892
X4  10.4481  1   0.001228 **
```

Example: disease outbreak

```
> glm4 = glm(Y~., data=mydata[,-1], family=binomial("logit")); summary(glm4)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.31293    0.64259  -3.599 0.000319 ***
X1           0.02975    0.01350   2.203 0.027577 *
X2           0.40879    0.59900   0.682 0.494954
X3          -0.30525    0.60413  -0.505 0.613362
X4           1.57475    0.50162   3.139 0.001693 **
---
Null deviance: 122.32  on 97  degrees of freedom
Residual deviance: 101.05  on 93  degrees of freedom
AIC: 111.05
>
> summary(glm(Y~X2+X3+X4,  family=binomial("logit")))

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4392     0.4690  -3.068  0.00215 **
X2            0.2351     0.5752   0.409  0.68278
X3           -0.4779     0.5829  -0.820  0.41230
X4            1.6203     0.4857   3.336  0.00085 ***
---
Null deviance: 122.32  on 97  degrees of freedom
Residual deviance: 106.20  on 94  degrees of freedom
AIC: 114.2
```

Example: disease outbreak

- $H_0 : \beta_4 = 0$. The P-value of this test is .001. We conclude that X_4 should not be dropped from the model.
- $H_0 : \beta_2 = \beta_3 = 0$.

```
> anova(glm(Y~X1+X4,  family=binomial("logit")), glm4, test="Chisq")
Analysis of Deviance Table

Model 1: Y ~ X1 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        95     102.26
2        93     101.05  2   1.2052   0.5474
```

  - The P-value suggests that socioeconomic status can be dropped from the model containing X_1 and X_4.
  - However, this variable was considered *a priori* to be important.
  - In addition, the estimated regression coefficients for X_1 and X_4 and their standard errors are not appreciably affected by whether or not socioeconomic status is in the regression model.
  - Hence, it was decided to keep socioeconomic status in the logistic regression model in view of its *a priori* importance.

Outline

Inference about Mean Response

- Let $\boldsymbol{X}_h = (1, X_{h1}, \ldots, X_{h,p-1})'$ denote the vector of explanatory variables.
- The corresponding mean response is

$$\pi_h = \frac{\exp(\boldsymbol{X}_h'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}_h'\boldsymbol{\beta})} = \left\{1 + \exp(-\boldsymbol{X}_h'\boldsymbol{\beta})\right\}^{-1}.$$

- Estimate π_h by

$$\hat{\pi}_h = \frac{\exp(\boldsymbol{X}_h'\hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{X}_h'\hat{\boldsymbol{\beta}})} = \left\{1 + \exp(-\boldsymbol{X}_h'\hat{\boldsymbol{\beta}})\right\}^{-1}.$$

Inference about Mean Response

- An approximate $(1-\alpha)$ confidence interval for $\boldsymbol{X}_h'\beta$ has lower and upper limits

  $$L = \boldsymbol{X}_h'\hat{\beta} - z_{1-\alpha/2}s\{\boldsymbol{X}_h'\hat{\beta}\}$$

  and

  $$U = \boldsymbol{X}_h'\hat{\beta} + z_{1-\alpha/2}s\{\boldsymbol{X}_h'\hat{\beta}\}$$

  where

  $$s^2\{\boldsymbol{X}_h'\hat{\beta}\} = \boldsymbol{X}_h'\boldsymbol{V}\{\hat{\beta}\}\boldsymbol{X}_h.$$

- An approximate $(1-\alpha)$ confidence interval for π_h has lower and upper limits

  $$L^* = \{1+\exp(-L)\}^{-1}$$

  and

  $$U^* = \{1+\exp(-U)\}^{-1}.$$

Example: disease outbreak

- Find an approximate 95% CI for the probability that persons 10 years old who are of lower socioeconomic status and live in sector 1 have contracted the disease.
- $\boldsymbol{X}_h = (1, 10, 0, 1, 0)'$.
- The mean response is $\pi_h = \left\{1 + \exp(-\boldsymbol{X}_h'\beta)\right\}^{-1}$.
- The point estimate of the logit mean response: $\boldsymbol{X}_h'\hat{\beta}$ = -2.32. Its standard error is $s\{\boldsymbol{X}_h'\hat{\beta}\}$ = .54.
- An approximate 95% CI for $\boldsymbol{X}_h'\beta$ is $(-3.38, -1.26)$.
- An approximate 95% CI for π_h is (.033, .22).

```
> predict(glm4, data.frame(X1=10, X2=0, X3=1, X4=0), se.fit=T, type="link")
$fit
1
-2.320688

$se.fit
[1] 0.5426989
```

Outline

Prediction of a New Observation

- For forecasting a binary response variable, predict the outcome to be 1 if $\hat{\pi}_h$ is large and 0 otherwise.
- How large is large?
- Different approaches to determining the cutoff point.
  1. Use 0.5 as the cutoff.

$$Y_{h(\text{new})} = \begin{cases} 1 & ; \quad \hat{\pi}_h > 0.5 \\ 0 & ; \quad \hat{\pi}_h \leq 0.5 \end{cases}$$

  2. Find the best cutoff based on data in the sense that the proportion of incorrect predictions is the lowest.
  3. Find the best cutoff that uses prior probabilities and costs of incorrect prediction.

Sensitivity and Specificity

- **Sensitivity** is the true positive rate

  $$\text{sensitivity}(c) = P(\hat{Y} = 1 | Y = 1) = \frac{\sum_{i=1}^{n} I(\hat{\pi}_i > c, Y_i = 1)}{\sum_{i=1}^{n} I(Y_i = 1)}$$

  where c is a given cutoff value.

- **Specificity** is the true negative rate

  $$\text{specificity}(c) = P(\hat{Y} = 0 | Y = 0) = \frac{\sum_{i=1}^{n} I(\hat{\pi}_i < c, Y_i = 0)}{\sum_{i=1}^{n} I(Y_i = 0)}$$

- Therefore $1 - \text{sensitivity}(c)$ is the false negative rate and $1 - \text{specificity}(c)$ is the false positive rate.

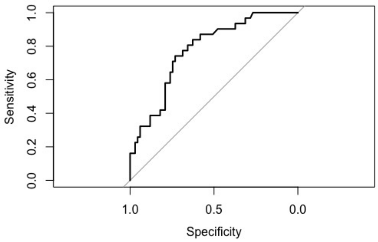
Receiver Operating Characteristic (ROC) Curve

- ROC curve plots sensitivity(c) as a function of $1-$ specificity(c) for all $c \in [0,1]$.
- **Area under the ROC curve** (AUC) estimates the probability that the predictions and the outcomes are concordant.
- General guidelines of interpreting area under the ROC curve:

AUC	Interpretation
≈ 0.5	Prediction is not better than random guess
$[0.7, 0.8]$	Acceptable discrimination
$[0.8, 0.9]$	Excellent discrimination
$[0.9, 1.0]$	Outstanding discrimination

Example: disease outbreak

```
> library(pROC)
> disease.roc = roc(Y ~ fitted(glm4))
> plot(disease.roc)
> auc(disease.roc)
Area under the curve: 0.7764
```



Outline

Model Selection Criteria

- For logistic regression, AIC and BIC are commonly-used criteria

$$\begin{aligned} \mathrm{AIC}_p &= -2\log\mathcal{L}(\hat{\beta}) + 2p \\ \mathrm{BIC}_p &= -2\log\mathcal{L}(\hat{\beta}) + p\log(n) \end{aligned}$$

- Promising models have relatively small values.
- The penalty terms are $2p$ for AIC and $p\log(n)$ for BIC.
- Most software packages also report $-2\log\mathcal{L}(\hat{\beta})$, which always increases as more explanatory variables are added to the model.

Model Selection

- The idea of best subsets in multiple linear regression applies here.
- A best subsets procedure identifies a group of subset models that give the best values of a given criterion.
- When the number of explanatory variables is large, however, all-possible best subsets may not be feasible.
- In this case, a stepwise selection procedure offers a feasible approach.
- The ideas of forward selection, backward elimination, and stepwise selection continue to apply.
- The rule for adding or deleting an explanatory variable often involves a p-value from the Wald test, AIC, BIC, etc.

Example: disease outbreak

Best Subsets procedure:

```
> library(leaps)
> source("myregsub.R")
> round(my.regsub(mydata[,2:5],Y, nbest=4, method="exhaustive",nvmax=4),3)
  (Intercept) X1 X2 X3 X4   rsq    rss adjr2     cp    bic
1           1  0  0  0  1 0.151 17.997 0.142  5.743 -6.852
1           1  1  0  0  0 0.077 19.555 0.068 14.373  1.280
1           1  0  0  1  0 0.040 20.349 0.030 18.774  5.181
1           1  0  1  0  0 0.015 20.874 0.005 21.685  7.679
2           1  1  0  0  1 0.199 16.971 0.182  2.057 -8.020
2           1  0  0  1  1 0.160 17.797 0.143  6.634 -3.362
2           1  0  1  0  1 0.157 17.871 0.139  7.041 -2.959
2           1  1  0  1  0 0.102 19.027 0.083 13.452  3.187
3           1  1  1  0  1 0.207 16.810 0.182  3.162 -4.372
3           1  1  0  1  1 0.204 16.865 0.179  3.465 -4.053
3           1  0  1  1  1 0.162 17.763 0.135  8.445  1.034
3           1  1  1  1  0 0.107 18.932 0.078 14.925  7.281
4           1  1  1  1  1 0.208 16.781 0.174  5.000  0.043
```

Example: disease outbreak

Forward Stepwise + BIC

```
glm0 = glm(Y~1, family=binomial("logit"))
step(glm0, scope=list(upper=glm4), direction="both", k=log(length(Y)) )
```

Step 1:

```
Start:  AIC=126.9
Y ~ 1

       Df Deviance    AIC
+ X4    1   107.53 116.70
+ X1    1   114.91 124.08
<none>      122.32 126.90
+ X3    1   118.23 127.40
+ X2    1   120.88 130.05
```

Step 2:

```
Step:  AIC=116.7
Y ~ X4

       Df Deviance    AIC
+ X1    1   102.26 116.01
<none>      107.53 116.70
+ X3    1   106.37 120.13
+ X2    1   106.88 120.64
- X4    1   122.32 126.90
```

Example: disease outbreak

Step 3:

```
Step:  AIC=116.01
Y ~ X4 + X1

       Df Deviance    AIC
<none>       102.26 116.01
- X1    1    107.53 116.70
+ X2    1    101.31 119.65
+ X3    1    101.52 119.86
- X4    1    114.91 124.08

Call:  glm(formula = Y ~ X4 + X1, family = binomial("logit"))

Coefficients:
(Intercept)           X4           X1
-2.33515      1.67345      0.02929

Degrees of Freedom: 97 Total (i.e. Null);  95 Residual
Null Deviance:     122.3
Residual Deviance: 102.3  AIC: 108.3
```

Example: disease outbreak

Try also:

```
##### Forward Stepwise + AIC
step(glm0, scope=list(upper=glm4), direction="both")

##### Forward selection + AIC
step(glm0, scope=list(upper=glm4), direction="forward")

##### Backward elimination + AIC
step(glm4)
```