Slide 1

Reference: Agresti, Sections 5.2-5.5

- The logistic regression model is defined as

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta},$$

- β_j represents the **change in log-odds** resulting from a one-unit increase in x_{ij}, holding all other covariates constant.
- The coefficients β_j's correspond to **log-odds ratios**.
- The logit transformation is one-to-one

$$\pi_i = \text{ logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}.$$

- This is an S-shaped function. As $\boldsymbol{x}_i$ increases, π_i increases when $\beta > 0$, and π_i decreases when $\beta < 0$.

Slide 2

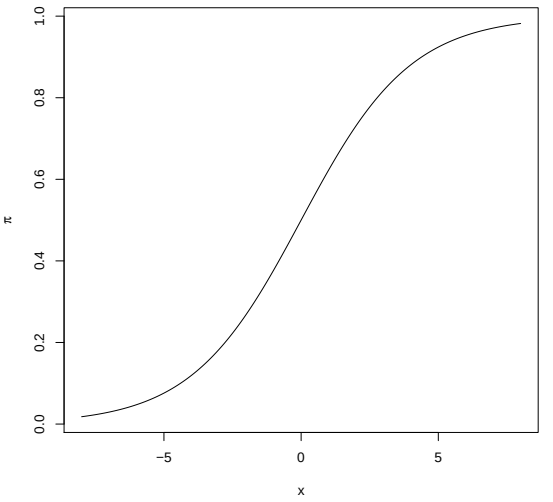

Figure 1: Logistic regression function, $\beta > 0$

Slide 3

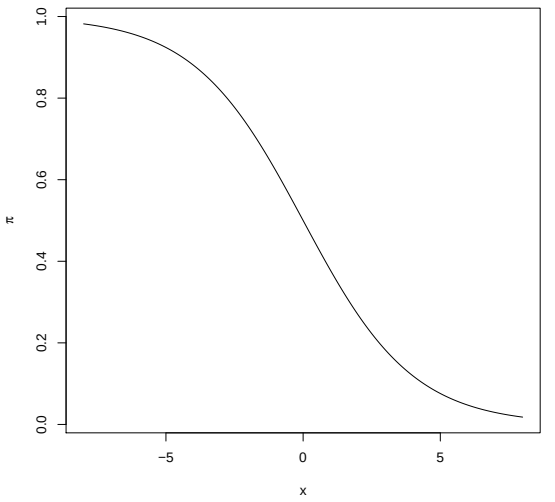

Figure 2: Logistic regression function, $\beta < 0$

Slide 4

Maximum likelihood estimation

- Since the logit is the canonical link for binomial data, Newton-Raphson and Fisher scoring are the same.
- We iterate until convergence

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right)^{-1} \boldsymbol{X}' \left(\boldsymbol{y} - \boldsymbol{\mu}^{(t)}\right),$$

where

$$\mu_i^{(t)} = n_i \pi_i^{(t)}, \quad \boldsymbol{W}^{(t)} = \mathrm{Diag}(n_i \pi_i^{(t)}(1 - \pi_i^{(t)})),$$

$$\pi_i^{(t)} = \frac{e^{\boldsymbol{x}_i' \boldsymbol{\beta}^{(t)}}}{1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}^{(t)}}}.$$

Slide 5

- We could also use the iterative re-weighted least squares approach

$$\boldsymbol{\beta}^{(t)} = \left(\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)},$$

  where the working dependent variable $\boldsymbol{z}^{(t)}$ has elements

$$z_i^{(t)} = \eta_i^{(t)} + \frac{y_i - \mu_i^{(t)}}{\mu_i^{(t)}(n_i - \mu_i^{(t)})} n_i.$$

- With large samples

$$\hat{\boldsymbol{\beta}} \xrightarrow{d} N(\boldsymbol{\beta}, (\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}).$$

Slide 6

Wald test

- We can test $H_0 : \beta_j = 0$ using the Wald test

$$z = \frac{\hat{\beta}_j}{\sqrt{\widehat{var}(\hat{\beta}_j)}} \xrightarrow{d} N(0,1),$$

  where $\widehat{var}(\hat{\beta}_j)$ is the (j,j)-th element of $\widehat{Var}(\hat{\boldsymbol{\beta}})$.

- The Wald test can be used to derive a confidence interval for β_j

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\widehat{var}(\hat{\beta}_j)},$$

- The Wald test can be applied to test hypotheses concerning several coefficients

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\widehat{Var}^{-1}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \chi^2_{\dim(\beta)}.$$

Slide 7

Likelihood ratio test

- Consider partitioning the model matrix and the vector of coefficients into two components

$$\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2) \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$$

  with p_1 and p_2 elements, respectively.

- Consider testing the hypothesis

$$H_0 : \boldsymbol{\beta}_2 = 0.$$

- Let $D(\boldsymbol{X}_1)$ and $D(\boldsymbol{X}_1 + \boldsymbol{X}_2)$ denote respectively the deviances of a model with $\boldsymbol{X}_1$ and a model with all variables in $\boldsymbol{X}$

$$D(\boldsymbol{X}_1) - D(\boldsymbol{X}_1 + \boldsymbol{X}_2) \xrightarrow{d} \chi^2_{p_2}$$

Slide 8

Inference on functions of $\boldsymbol{\beta}$

- By the delta method, we have

$$g(\hat{\boldsymbol{\beta}}) \sim N\left(g(\boldsymbol{\beta}), \left(\frac{\partial g(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}}\right)^T \widehat{Var}(\hat{\boldsymbol{\beta}}) \left(\frac{\partial g(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}}\right)\right).$$

- One important non-linear function is π

$$\pi_i = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}.$$

Slide 9

Fitted values

- Verify that an estimated variance for $\hat{\pi}_i$ is

$$\widehat{Var}(\hat{\pi}_i) = \hat{\pi}_i^2(1-\hat{\pi}_i)^2 \boldsymbol{x}_i' \left(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i.$$

- Define $\boldsymbol{U} = \text{Diag}(\pi_i(1-\pi_i))$, then

$$\widehat{Var}(\hat{\boldsymbol{\pi}}) = \boldsymbol{U}\boldsymbol{X}\left(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{U}.$$

- Since this estimate borrows information across the entire dataset $y_1, \ldots, y_N$, it should be smaller than the estimated variance of $p_i = y_i/n_i$, $\widehat{Var}(p_i) = p_i(1-p_i)/n_i$, provided the model is true.

Slide 10

Testing goodness-of-fit

- In logistic regression, the fitted values are

$$\hat{\mu}_i = n_i\hat{\pi}_i = n_i\frac{\exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}, \quad i = 1, \ldots, N.$$

- The deviance G^2 and the Pearson $\mathcal{X}^2$ compare the fitted values $\hat{\mu}_i$'s to the y_i's.

- That is, these statistics test the fit of the logit model (a p-parameter model) against the saturated model that allows π_i to lie anywhere in $(0, 1)$ (an N-parameter model).

Slide 11

Deviance test statistic

- The deviance test statistic is

$$G^2 = 2\sum_{i=1}^{N}\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i\hat{\mu}_i}\right)\right\} \xrightarrow{d} \chi^2_{N-p}.$$

- As the group sizes $n_i \to \infty$, $\forall i$, the deviance statistic converges to χ^2_{N-p}, where N is the number of groups and p is the number of parameters in the model (including the intercept).
- With individual data, the distribution of the deviance does not converge to a chi-squared and it can not be used as a goodness-of-fit test.

Slide 12

Pearson goodness-of-fit statistic

- The **Pearson residuals** are defined as

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{Var}(\hat{\mu}_i)}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

- The **Pearson goodness-of-fit statistic** is

$$\mathcal{X}^2 = \sum_{i}^{N} e_i^2 \xrightarrow{d} \chi^2_{N-p}.$$

- With grouped data and large samples, the Pearson $\mathcal{X}^2$ statistic has a χ^2_{N-p} distribution and is equivalent to the deviance.
- It cannot be used as a goodness of fit test with individual data.

Slide 13

Let's revisit the Beetles data.

```
Beetles$alive = Beetles$n - Beetles$dead
attach(Beetles)

fit.logit = glm(cbind(dead, alive) ~ logdose, family=binomial(link=logit))
summary(fit.logit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.740      5.182  -11.72   <2e-16 ***
logdose       34.286      2.913   11.77   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.116  on 6  degrees of freedom
```

Slide 14

Goodness-of-fit statistics

```
# Deviance test statistic
Residual deviance:  11.116  on 6  degrees of freedom

# Pearson chi-squared goodness-of-fit, df=6
sum(resid(fit.logit, type="pearson")^2)
[1] 9.906715

# p-values
1-pchisq(11.116, 6);  1-pchisq(9.906715, 6)
[1] 0.0848568
[1] 0.1286358
```

The deviance $G^2 = 11.12$ and Pearson $\chi^2 = 9.91$ have $df = 8 - 2 = 6$ and show slight evidence of lack-of-fit (p-values ≈ 0.1).

Slide 15

A 0.1 unit increase in $\log_{10}$-concentration

- increases the log-odds of death by 3.4286
- is associated with a log-odds ratio of 3.4286
- multiplies the odds of death by $\exp(3.4286) = 30.83$.

Slide 16

Obtaining fitted values and residuals

```
pihat.logit = fit.logit$fitted.values          # fitted values
pearson.res = resid(fit.logit, type="pearson")         # Pearson residuals

> cbind(logdose, dead/n, pihat.logit, pearson.res)
  logdose           pihat.logit pearson.res
1   1.691 0.1016949  0.05937747   1.3753932
2   1.724 0.2166667  0.16366723   1.1096257
3   1.755 0.2903226  0.36162283  -1.1684774
4   1.784 0.5000000  0.60490961  -1.6058900
5   1.811 0.8253968  0.79440490   0.6086835
6   1.837 0.8983051  0.90405532  -0.1499696
7   1.861 0.9838710  0.95546748   1.0842287
8   1.884 1.0000000  0.97925643   1.1273769
```

Slide 17

Hosmer-Lemeshow goodness-of-fit statistic

- Hosmer and Lemeshow (1989) have proposed a procedure that can be used with individual data even if there are no common covariate patterns.
- Hosmer-Lemeshow recommend forming ten groups with predicted probabilities $(0 - 0.1], (0.1 - 0.2], \ldots, (0.9, 1)$.
- Other ways of pooling the data can be considered.
- One can then compute expected counts for each group and compare them with the observed values

$$\chi^2_{HL} = \sum_{i=1}^{G} \frac{[(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij})^2}{(\sum_j \hat{\pi}_{ij})[1 - (\sum_j \hat{\pi}_{ij})/n_i]}$$

y_{ij} = binary outcome for observation j in group i

$\hat{\pi}_{ij}$ = fitted probability for ungrouped data.

Slide 18

Hosmer-Lemeshow goodness-of-fit statistic

- The statistic does not have a chi-squared distribution, but simulation studies have shown that in large samples it is similar to a chi-squared distribution with $df = G - 2$.
- A large value of the Hosmer-Lemeshow statistic indicates some lack of fit but does not provide insight about its nature.

Slide 19

Example: Low birth weight (saved as `BirthWeight.txt`)

ID	Identification code
LOW	Low birth weight (0=weight>2500g, 1=weight<2500g)
AGE	Age of mother in years
LWT	Weight in pounds at last menstrual period
RACE	Race (1=White, 2=Black, 3=Other)
SMOKE	Smoking status during pregnancy (1=Yes, 0=No)
PTL	History of premature labor (0=None, 1=One, 2=Two, etc.)
HT	History of hypertension (1=Yes, 0=No)
UI	Presence of uterine irritability (1=Yes, 0=No)
FTV	Number of physician visits during the first trimester
BWT	Birth weight in grams

Slide 20

- We want to model the risk of low birthweight (`low`=1) in terms of maternal pre-pregnancy weight (`lwt`), race, smoking status, history of premature labor (`ptl`) and history of hypertension (`ht`).
- Very few observations share the same `lwt` value, so there will be one observation in most of the cells created by covariate patterns.

Slide 21

```
attach(BrithWeight)
bwt.logit = glm(low ~ lwt+smoke+as.factor(race)+ptl+ht, family="binomial")

library(ResourceSelection)
res = hoslem.test(bwt.logit$y, fitted(bwt.logit))
```

Slide 22

```
> res
Hosmer and Lemeshow goodness of fit (GOF) test
data:  bwt.logit$y, fitted(bwt.logit)
X-squared = 7.9557, df = 8, p-value = 0.4378

> cbind(res$observed, res$expected)
              y0 y1     yhat0     yhat1
[0.038,0.104] 18  1 17.611209  1.388791
(0.104,0.14]  16  3 16.739105  2.260895
(0.14,0.199]  14  5 15.816089  3.183911
(0.199,0.261] 16  3 14.473849  4.526151
(0.261,0.284] 18  2 14.474378  5.525622
(0.284,0.322]  9  8 11.795959  5.204041
(0.322,0.361] 12  7 12.549143  6.450857
(0.361,0.442] 12  7 11.610390  7.389610
(0.442,0.603] 10  9  9.202104  9.797896
(0.603,0.837]  5 14  5.727773 13.272227
```

Thus, the model appears to be adequate for these data, so we can go ahead and look at the parameter estimates:

```
> summary(bwt.logit)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.117888   0.944152   0.125   0.9006
lwt              -0.016580   0.006857  -2.418   0.0156 *
smoke             0.946179   0.394947   2.396   0.0166 *
as.factor(race)2  1.290381   0.522377   2.470   0.0135 *
as.factor(race)3  0.910325   0.428269   2.126   0.0335 *
ptl               0.602481   0.335233   1.797   0.0723 .
ht                1.745050   0.694902   2.511   0.0120 *

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 204.90  on 182  degrees of freedom
```

Slide 23

Is grouping a good idea?

- If the model is true, adding the y_i's causes no loss of information.
- If the model is not true (e.g. important covariates omitted), then two models with same $\boldsymbol{x}_i$'s may have different π_i's.
- Grouping may sacrifice our ability to detect departures from the model.

Slide 24