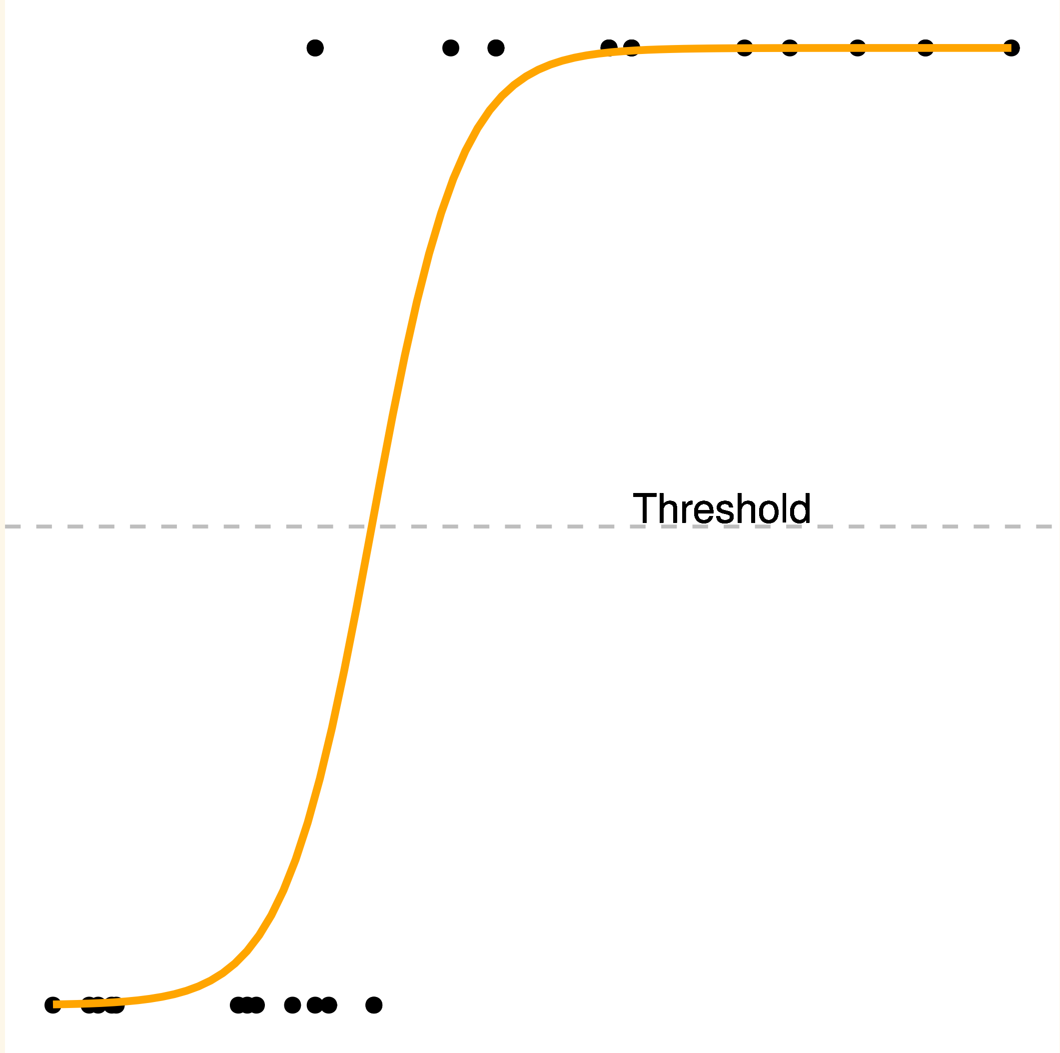
Logistic regression for binary responses: Inference

Stat 230

May 13 2022

Overview



Today:

logistic regression model

- inference

Deviance

- model comparisons

The logistic model

- Our Bernoulli responses are modeled as a function of predictors $X_i = x_{1,i}, \dots, x_{p,i}$ through the probability of success:

$$Y_i \mid X_i \overset{\text{indep.}}{\sim} \text{Bern}(\pi\left(X_i\right))$$

- Log odds of success (logit):

$$\eta_i = \log\left(\frac{\pi\left(X_i\right)}{1-\pi\left(X_i\right)}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}$$

- Probability of success:

$$\pi\left(X_i\right) = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_p x_{p,i}}}{1+e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_p x_{p,i}}}$$

Generalized linear model

The **kernel mean function** defines the expected value (mean) of Y as a function of η.

- in a logistic model, the kernel mean function is the logistic function $E(Y \mid X) = \pi(X) = \frac{e^\eta}{1+e^\eta}$

The **link function** defines the linear combination η as a function of the mean of Y.

- in a logistic model, the link function is the logit function $\eta = \log(\pi/(1-\pi))$
- These two functions are inverses of one another.

MLR vs logistic inference comparison

MLR

- Estimation: Maximum likelihood
- One β inference: t-distribution inference
- Model comparison inference: ANOVA F-tests

Logistic regression

- Estimation: Maximum likelihood
- One β inference: z-distribution inference
- Model comparison inference: Drop-in-deviance Chi-square tests

Inference and estimation

- Estimation done using maximum likelihood estimation (MLE)
- Likelihood is the probability of the observed data, written as a function of our unknown β 's

$$L(\beta; data) = \prod_{i=1}^{n} \pi(X_i)^{y_i}(1 - \pi\left(X_i\right))^{1-y_i}$$

- Find the β' s that maximize $L(\beta;$ data)
- Unlike SLR or MLR, there is no "closed form" for these MLE $\hat{\beta}_i$
- Software uses a numerical optimization method to compute the MLEs $\hat{\beta}_i$ and the standard errors $SE\left(\hat{\beta}_i\right)$

Inference and estimation

- MLE estimates of $\hat{\beta}_i$ are approximately normally distributed and unbiased when n is "large enough."

Confidence intervals for one β_i

- A $C\%$ confidence interval for β_i equals

$$\hat{\beta}_i \pm z^* SE\left(\hat{\beta}_i\right)$$

where z^* is the $(100 - C)/2$ percentile from the $N(0,1)$ distribution.

Inference and estimation

Hypothesis tests for one β_i

- The usual test results given by standard regression output tests whether a parameter value (intercept or slope) is equal to 0 vs. not equal to 0 :

$$H_0 : \beta_i = 0 \quad H_A : \beta_i \neq 0$$

with a test stat of

$$z = \frac{\hat{\beta}_i - 0}{SE\left(\hat{\beta}_i\right)}$$

The $N(0, 1)$ is used to compute the p-value that is appropriate for whatever H_A is specified.

Inference and estimation

Drop-in-deviance Model comparison tests

- In a GLM, deviance is measures something similar to residual sum of squares
- When the GLM = MLR, deviance is the same as SSR.
- We use G^2 to denote deviance of a model
- Our model comparison test compares G^2 from two competing models

Drop in Deviance test

(1) **Hypotheses:**

$$\begin{aligned} H_0 &: \text{ reduced model} \\ H_A &: \text{ full model} \end{aligned}$$

(2) **Test Statistic:** The likelihood ratio test (LRT) stat compares the drop in deviance from the reduced to the full models

$$LRT = G^2_{\text{reduced}} - G^2_{\text{full}}$$

(3) When n is "large enough", the LRT will have a chi-square (χ^2) distribution with $df = df_{\text{reduced}} - df_{\text{full}} = \#$ terms tested.

The p-value is a right tailed area

$$\text{p-value} = P\left(\chi^2 > LRT\right) = 1 - pchisq(LRT, df)$$

Drop in Deviance test

Special cases of drop in deviance tests:

- The overall drop in deviance test compares a null "intercept only" model to a logistic model:

$$\begin{aligned} H_0 &: \log(\text{odds}) = \beta_0 \\ H_A &: \log(\text{odds}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \end{aligned}$$

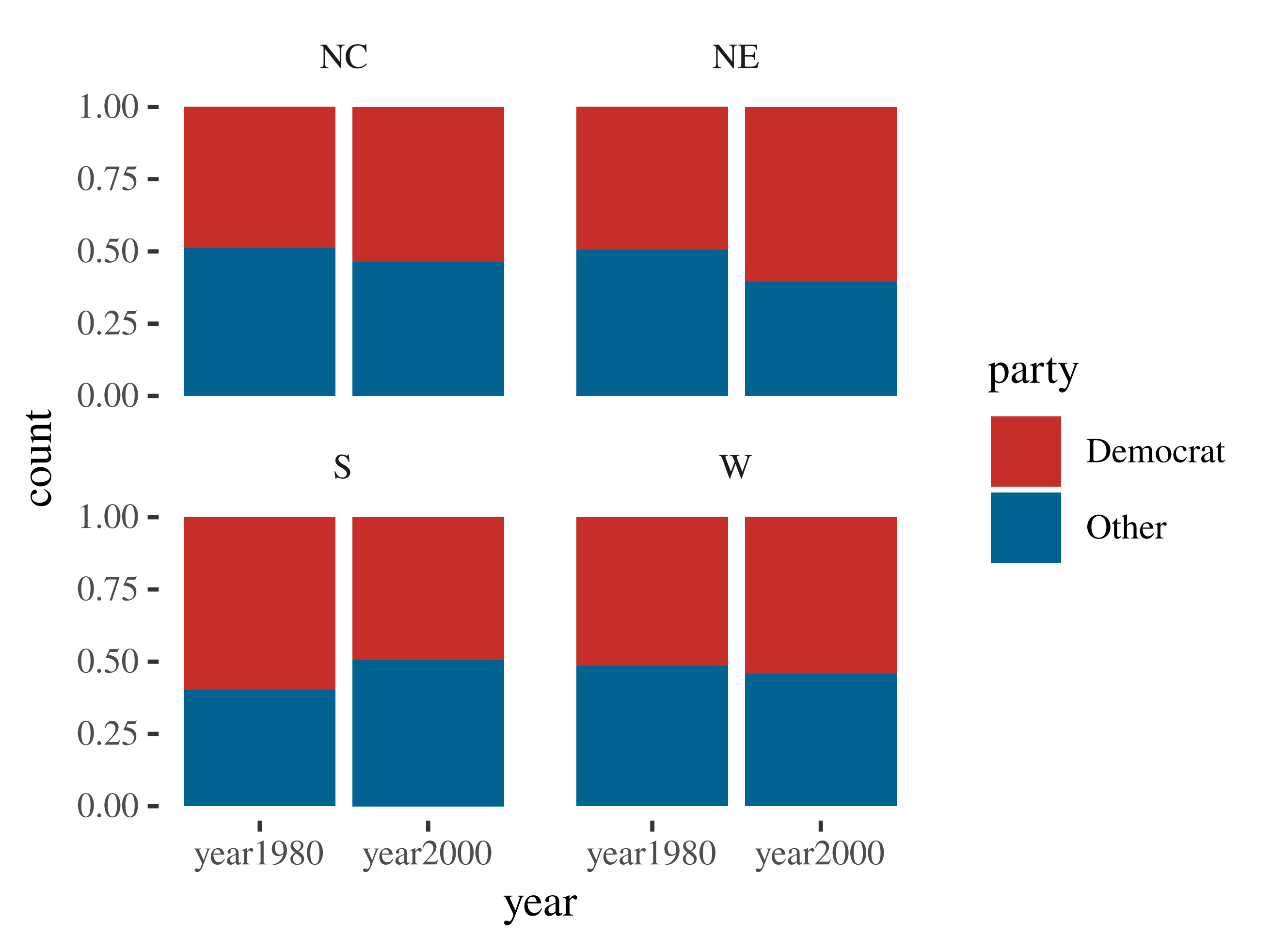
- Null deviance is similar in spirit to the total sum of squares in ANOVA.

If our reduced and full models differ by one term, then the drop in deviance test will test the same hypotheses as the z-test (a.k.a. Wald test) for the term,

- but the two methods of testing are not identical.
- tests will usually agree, but if they do not, use the drop in deviance LRT test results.

How does temporal changes in party ID differ across regions?

```
library(dplyr)
library(ggthemes)
nes$party <- recode_factor(nes$dem,
                           `1`="Democrat",
                           `0`="Other")
ggplot(nes, aes(x=year, fill = party)) +
 geom_bar(position="fill") +
 facet_wrap(~region) +
scale_fill_wsj()
```



Example: NES

$$\text{odds}(\text{dem} \mid x) = e^{\beta_0 + \beta_{NE} NE + \beta_S S + \beta_W W + \beta_{2000} \text{ year 2000} + \beta_{NE:2000} NE\text{: year 2000} + \beta_{S:2000} S\text{: year 2000} + \beta_{W:2000} W\text{: year 2000}}$$

Interpretation in terms of odds

```
nes_glm1 <- glm(dem ~ region*year , data=nes, family = binomial)
tidy(nes_glm1, conf.int=TRUE, exponentiate=TRUE)
```

```
# A tibble: 8 × 7
  term                   estimate std.error statistic p.value conf.low conf.high
  <chr>                     <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 (Intercept)               0.955     0.124    -0.371 0.711      0.749      1.22
2 regionNE                  1.03      0.182     0.155 0.877      0.720      1.47
3 regionS                   1.57      0.162     2.79  0.00534    1.14       2.16
4 regionW                   1.12      0.192     0.575 0.565      0.767      1.63
5 yearyear2000              1.22      0.169     1.18  0.240      0.876      1.70
6 regionNE:yearyear2000     1.29      0.259     0.977 0.328      0.776      2.14
7 regionS:yearyear2000      0.531     0.221    -2.86  0.00427    0.344      0.819
8 regionW:yearyear2000      0.917     0.257    -0.338 0.735      0.553      1.52
```

Example: NES

$$\text{logit}(\text{dem} \mid x) = \beta_0 + \beta_{NE} NE + \beta_S S + \beta_W W + \beta_{2000}\ \text{year 2000}$$
$$+ \beta_{NE:2000} NE:\ \text{year 2000} + \beta_{S:2000} S:\ \text{year 2000} + \beta_{W:2000} W:\ \text{year 2000}$$

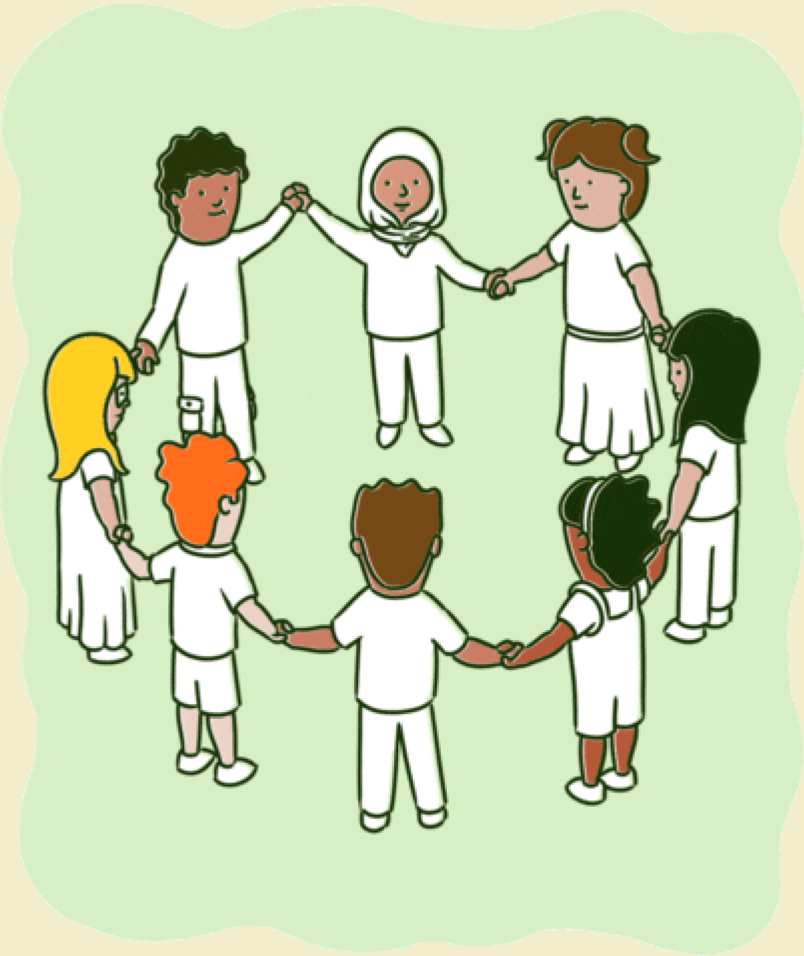
Interpretation in terms of log of odds

```
nes_glm1 <- glm(dem ~ region*year , data=nes, family = binomial)
tidy(nes_glm1, conf.int=TRUE)
```

```
# A tibble: 8 × 7
  term                   estimate std.error statistic p.value conf.low conf.high
  <chr>                     <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 (Intercept)             -0.0458     0.124    -0.371 0.711     -0.289     0.197
2 regionNE                 0.0281     0.182     0.155 0.877     -0.328     0.384
3 regionS                  0.451      0.162     2.79  0.00534    0.134     0.770
4 regionW                  0.110      0.192     0.575 0.565     -0.266     0.487
5 yearyear2000             0.199      0.169     1.18  0.240     -0.133     0.531
6 regionNE:yearyear2000    0.253      0.259     0.977 0.328     -0.254     0.762
7 regionS:yearyear2000    -0.633      0.221    -2.86  0.00427   -1.07     -0.199
8 regionW:yearyear2000    -0.0870     0.257    -0.338 0.735     -0.592     0.418
```

Your Turn 1

05:00



- Go over to the in class activity file
- Go over the class activity in your group

Example: NES

1980 (baseline year): The difference in log-odds between the S region and the NC (baseline) region is β_{south} :

$$\text{logit}(\text{ region } = S, \text{ year } = 1980) - \text{logit}(\text{ region } = NC, \text{ year } = 1980) = \beta_{\text{south}}$$

1980 (baseline year): In 1980 , the odds of being a Democrat in the south was 1.57 times the odds in the north central region. (part f)

$$\frac{\widehat{\text{odds}(\text{ region } = S, \text{ year } = 1980)}}{\text{odds}(\text{ region } = NC, \text{ year } = 1980)} = e^{\hat{\beta}_{\text{south}}} = e^{0.451} = 1.57$$

- This effect is statistically significant $(\text{z} = 2.79, \text{p} = 0.00534)$.

$$H_0 : \beta_S = 0 \quad \text{test stat: } z = \frac{0.451 - 0}{0.162} = 2.79$$
$$p - \text{ value } = 2 \times P(Z < -2.79) = 2 \times \text{pnorm}(-2.79) = 0.00534$$

Example: NES

- region = S: The odds of being a Democrat in 2000 is 35% lower than being a Dem in 1980 in the South region. (part h)

$$\widehat{OR_{\text{South}}}\,(2000 \text{ vs.}1980\ ) = e^{\hat{\beta}_{\text{year 2000}}}\, e^{\hat{\beta}_{\text{year2000:South}}\,(1)} = e^{0.199} e^{-0.633} = 0.648$$

Compare OR in S vs. NC: $e^{\hat{\beta}_{2000:S\text{ suth}}} = 0.531$ is the factor change between the odds ratio for the South compared to NC regions:

$$\frac{OR_{\text{South}}\,\widehat{(2000 \text{ vs }.1980)}}{OR_{NC}(2000 \text{ vs. } 1980\ )} = e^{\hat{\beta}_{\text{2000: South}}} = e^{-0.633} = 0.531$$

- This is a statistically significant change $(\mathrm{z} = -2.86, \mathrm{p} = 0.00427\ )$

Deviance in GLMs

(Residual) Deviance is the term used to measure "unexplained" variation in the response.

- In MLR: deviance $=$ SSR

Small deviance:

- predicted $\hat{\pi}\left(X_i\right)$ are close to 1 when $y_i = 1$
- predicted $\hat{\pi}\left(X_i\right)$ are close to 0 when $y_i = 0$

Deviance will decrease as model terms are added.

Deviance

Logistic GLM deviance is the difference of two likelihoods

$$
\begin{aligned}
G^2 &= 2[\ln L(\bar{\pi}) - \ln L(\hat{\pi}(X))] \\
&= 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\pi}\left(X_i\right)}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{\pi}\left(X_i\right)}\right)\right]
\end{aligned}
$$

$L(\hat{\pi}(X))$: likelihood of the data that plugs in estimates $\hat{\pi}\left(X_i\right)$ from the logistic model.

$L(\bar{\pi})$: likelihood of the data that plugs in estimates $\bar{\pi} = y_i$, basing a case's "predicted" value only on the response observed for that case.

- called a saturated model
- will always have a higher likelihood than the logistic model:

$$L(\bar{\pi}) \geq L(\hat{\pi}(X))$$

Deviance and model comparison in R

```
anova(my.glm)
```

- gives the extra deviance explained by a term when it is added to the model above it in the table

```
anova(reduced.glm, full.glm, test = "Chisq")
```

- gives drop in deviance test results

Example: NES

```
anova(nes_glm1)
Analysis of Deviance Table

Model: binomial, link: logit

Response: dem

Terms added sequentially (first to last)

            Df Deviance Resid. Df Resid. Dev
NULL                         2231     3083.3
region       3   1.4306      2228     3081.9
year         1   0.0006      2227     3081.9
region:year  3  16.3611      2224     3065.5
```

Null deviance (no predictors): 3083.3

Deviance for region model: 3081.9

- adding region drops deviance by 1.4306

Example: NES

```
anova(nes_glm1)
Analysis of Deviance Table

Model: binomial, link: logit

Response: dem

Terms added sequentially (first to last)

            Df Deviance Resid. Df Resid. Dev
NULL                         2231     3083.3
region       3   1.4306      2228     3081.9
year         1   0.0006      2227     3081.9
region:year  3  16.3611      2224     3065.5
```

Deviance for region and year model: 3081.9

- adding year drops deviance by 0.0006

Deviance for region, year, region:year model: 3065.5

- adding region: year drops deviance by 16.3611

Example: NES

Does the effect of year on odds of being a Democrat depend on region?

$$
\begin{aligned}
H_0 &: \log(\text{ odds }) = \beta_0 + \beta_1 NE + \beta_2 S + \beta_3 W + \beta_4 Year2000 \\
H_A &: \log(\text{ odds }) = \beta_0 + \beta_1 NE + \beta_2 S + \beta_3 W + \beta_4 Y\text{ ear } 2000 + \beta_5 NE : 2000 \\
&\quad + \beta_6 S : 2000 + \beta_7 W : 2000
\end{aligned}
$$

```
nes_glm_red <- glm(dem ~ region+year , data=nes, family = binomial) # null model
tidy(nes_glm_red)
# A tibble: 5 × 5
  term         estimate std.error statistic p.value
  <chr>           <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)  0.0593     0.0958    0.619    0.536
2 regionNE     0.132      0.129     1.03     0.304
3 regionS      0.114      0.110     1.03     0.301
4 regionW      0.0681     0.128     0.533    0.594
5 yearyear2000 0.00203    0.0852    0.0238   0.981
```

Example: NES

```
anova(nes_glm_red, nes_glm1, test = "Chisq")
Analysis of Deviance Table

Model 1: dem ~ region + year
Model 2: dem ~ region * year
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      2227     3081.9
2      2224     3065.5  3   16.361 0.0009562 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The LRT stat equals $LRT = 3081.9 - 3065.5 = 16.361$
- degrees of freedom is 3 , so the p-value is

$$P\left(\chi^2 > 16.361\right) = 1 - pchisq(16.361, 3) = 0.00096$$

- We can conclude that the full model is better than the smaller model. There is at least one region's change in party affiliation between 1980 and 2000 that is different from the other regions.