Logistic regression and prediction

Agenda

- Exam 1
  - Wednesday September 21, in class
  - Covers material up through today (inclusive)
  - Closed notes
  - Bring a calculator (cannot use phone or laptop)
  - I won't ask you to write R code, but you may need to interpret R output
  - Questions similar to assignments and class activities
- Today: more logistic regression

Data

Data on 5720 Vietnamese children, admitted to hospital with possible dengue fever. Variables include:

- `Dengue`: whether the patient actually has dengue fever, based on a lab test (0 = no, 1 = yes)
- `Temperature`: patient's body temperature (in Celsius)
- `Abdominal`: whether the patient has abdominal pain (0 = no, 1 = yes)
- `HCT`: patient's hematocrit (proportion of red blood cells)
- `Age`: patient's age (in years)
- `Sex`: patient's sex
- + several others

Last time

$$Y_i \sim Bernoulli(\pi_i)$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\, Temperature_i + \beta_2\, Abdominal_i + \beta_3\, Temperature_i \cdot Abdominal_i$$

> Does the model improve when we add hematocrit (the proportion of red blood cells)?

Model

$$Y_i \sim Bernoulli(\pi_i)$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\, Temperature_i + \beta_2\, Abdominal_i$$
$$+ \beta_3\, Temperature_i \cdot Abdominal_i$$
$$+ \beta_4\, HCT_i$$

want to check relationship between

- HCT and Dengue (empirical log odds)
- Do empirical log odds plots to check for interactions w/ Temp & Abdominal

Lab3: involves checking these interactions

Class activity, Part I

https://sta214-f22.github.io/class_activities/ca_lecture_11.html

Class activity

> What is the estimated change in odds associated with a 1 point increase in hematocrit, holding temperature and abdominal pain constant?

Increases by a factor of $e^{\hat{\beta}_4} = e^{0.115} \approx 1.12$

Class activity

> How does the deviance change when we add hematocrit to the model?

Deviance w/out HCT: 6914

w HCT: 6744.8

=> Decrease by 169.2

Estimate Std Error z P(>|z|)

two-sided p-value

Class activity

(with those variables in the model)

> Researchers want to test whether there is a relationship between hematocrit and the probability a patient has dengue, after accounting for temperature and abdominal pain. Carry out a hypothesis test to investigate this research question.

$H_0: \beta_4 = 0$ $H_A: \beta_4 \neq 0$

LRT

$G = 169.2$

under H_0, $G \sim \chi^2_1$ ← # parameters tested

pchisq(169.2, df=1, lower.tail=F)

≈ 0

Wald

$z = \frac{12.753}{0.115} = \frac{0.115}{0.009}$

p-value ≈ 0

Comparing models

If deviance always decreases when I add additional variables, how can I assess whether including hematocrit substantially improves the model?

Option 1: Likelihood ratio test

- Is the change in deviance bigger than we would expect if hematocrit doesn't really matter?

Option 2: AIC (Akaike's Information Criterion)

AIC

> In linear regression, what quantity did we use to compare models with different numbers of parameters?

AIC

$R^2 = 1 - \frac{SSE}{SSTotal}$

more parameters $\Rightarrow$ SSE $\downarrow$ $\Rightarrow$ $R^2 \uparrow$

$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SSTotal/(n-1)}$

p = # parameters in model

> In linear regression, what quantity did we use to compare models with different numbers of parameters?

Adjusted R^2

- We can use something similar for logistic regression, called the *Akaike information criterion* (AIC)
- Motivation: penalize the deviance based on the number of parameters

Logistic regression: maximize likelihood
$\Longleftrightarrow$
minimize deviance (like SSE)

AIC

AIC: Suppose our model has p parameters (including the intercept). Then the AIC is

$$AIC = 2p + \text{deviance}$$

#parameters

want deviance to be small

AIC

Model 1: (adding hematocrit) $p = 5$

$6745 + 2(5)$

```
## Null Deviance:         6956
## Residual Deviance: 6745      AIC: 6755
```

Model 2: (no hematocrit) $p = 4$

$6914 + 2(4)$

```
## Null Deviance:         6956
## Residual Deviance: 6914      AIC: 6922
```

> Which model do we prefer, based on AIC?

Model 1! (smaller AIC)

Model comparison

Does the model improve when we add hematocrit (the proportion of red blood cells)?

- **Likelihood ratio test:** p-value ≈ 0
- **AIC:** AIC is smaller when we add hematocrit

Conclusion: We have convincing evidence that adding hematocrit improves the model.

A new question...

You report your results to the hospital, and they ask a follow-up question:

How good is your model at predicting whether a patient has dengue?

Making predictions

- For each patient in the data, we calculate $\widehat{\pi}_i$
- But, we want to decide which patients to treat. So we need to guess whether patient i has dengue $(Y_i = 1)$ or doesn't $(Y_i = 0)$

> How can we turn $\widehat{\pi}_i$ into a dengue prediction?

If $\hat{\pi}_i = 0$, guess $Y_i = 0$ $\Rightarrow$ $\hat{Y}_i = 0$

$\hat{\pi}_i = 1$ $\quad$ $\hat{Y}_i = 1$

$\hat{\pi}_i = 0.3$ $\quad$ $\hat{Y}_i = 0$

$$\hat{Y}_i = \begin{cases} 1 & \hat{\pi}_i \geq 0.5 \\ 0 & \hat{\pi}_i < 0.5 \end{cases}$$

$\leftarrow$ threshold

(could use other thresholds too)

Confusion matrix

3957 :
the patients who we correctly predicted do not have dengue

		Actual	
		$Y = 0$	$Y = 1$
Predicted	$\widehat{Y} = 0$	3957	1631
	$\widehat{Y} = 1$	66	66

bad "
good
patients we correctly predicted do have dengue

- For 3957 patients, we correctly predicted they did not have dengue
- For 66 patients, we correctly predicted they had dengue
- For 1631 patients, we incorrectly predicted they did not have dengue

> Did we do a good job at predicting?

Accuracy

Also look at performance within each group
$Y=0$ and $Y=1$

$Y=0$: $\frac{3957}{3957+66}$

$Y=1$: $\frac{66}{1631+66}$

		Actual	
		$Y=0$	$Y=1$
Predicted	$\widehat{Y}=0$	3957	1631
	$\widehat{Y}=1$	66	66

$$
\begin{aligned}
\text{Accuracy} &= \frac{\text{number of correct predictions}}{\text{number of observations}} \\
&= \frac{3957+66}{5720} \\
&= 0.703
\end{aligned}
$$

We correctly predict dengue status 70% of the time.

If we have unbalanced data, (lots more 0s or 1s)
Accuracy can be misleading

Class activity, Part II

https://sta214-f22.github.io/class_activities/ca_lecture_11.html

Class activity

		Actual	
		$Y=0$	$Y=1$
Predicted	$\widehat{Y}=0$	3990	503
	$\widehat{Y}=1$	33	1194

> What is the accuracy of the rapid test?

Class activity

> Which method would you prefer -- our logistic regression model, or the rapid test?