Prediction and hypothesis testing

Data

Question: What is the relationship between age and contraceptive use for women in Indonesia?

Data: 1473 Indonesian couples, with variables

- Y_i = contraceptive method used (1 = no use, 2 = long-term, 3 = short-term)
- X_i = Wife's age (numeric)

Last time: Fitted model

$$\log\left(\frac{\widehat{\pi}_{i(Short)}}{\widehat{\pi}_{i(None)}}\right) = -8.234 + 0.456 Age_i - 0.0065 Age_i^2$$

$$\log\left(\frac{\widehat{\pi}_{i(Long)}}{\widehat{\pi}_{i(None)}}\right) = -5.083 + 0.366 Age_i - 0.00628 Age_i^2$$

Last time: Predictions

```
##             Actual
## Prediction None Short Long
##       None   342   166  189
##       Short    0     0    0
##       Long   287   167  322
```

> How good are our predictions?

Really bad for short term use

ok for None & Long term use

Accuracy: $\frac{342 + 322}{1473} = 0.45$

Last time: Predictions

```
##             Actual
## Prediction None Short Long
##       None   342   166  189
##       Short    0     0    0
##       Long   287   167  322
```

We can also assess our predictions by comparing to random guessing.

> What are our predicted probabilities for each observation from random guessing?

Option 1: $\frac{1}{3}$ probability for each level (None, Short, Long)

Option 2: $P(\text{None}) = \frac{\#\text{ None}}{n}$, $P(\text{Short}) = \frac{\#\text{ Short}}{n}$, $P(\text{Long}) = \frac{\#\text{ Long}}{n}$

Random guessing

- If we don't have any data, our estimated probability would be 1/3 for each level
- If we have data but we don't use age, our estimated probability for each level is just the proportion of observations in that group:

```
table(cmc_data$Choice)/nrow(cmc_data)
```

```
##
##      None     Short      Long
## 0.4270197 0.2260692 0.3469111
```

Class activity

https://sta214-f22.github.io/class_activities/ca_lecture_16.html

Class activity

What would our confusion matrix look like if our predictions randomly assigned each person to one of the three categories, with a 1/3 chance for each category?

Y

$\hat{Y}$		None	Short	Long
	N	210	111	170
	S	210	111	170
	L	209	111	171
		629	333	511

$\frac{629}{3} \approx 210$ $\quad \frac{333}{3} \approx 111$ $\quad \frac{511}{3} \approx 170$

Class activity

> What would our confusion matrix look like if our predictions randomly assigned each person to one of the three categories, with a 1/3 chance for each category?

Something like

		Actual		
		None	Short	Long
Predicted	None	210	111	170
	Short	210	111	170
	Long	209	111	171

Class activity

		Actual		
		None	Short	Long
Predicted	None	210	111	170
	Short	210	111	170
	Long	209	111	171

> What is the accuracy of our predictions in this confusion matrix?

Abat $\frac{1}{3}$

Class activity

What would our confusion matrix look like if for every individual, we just predicted the most common contraception choice in the data?

Class activity

> What would our confusion matrix look like if for every individual, we just predicted the most common contraception choice in the data?

The most common choice is None, so

		Actual		
		None	Short	Long
Predicted	None	629	333	511
	Short	0	0	0
	Long	0	0	0

factor(Choice, levels = c("None",
"Short",
"Long"))

Class activity

		Actual		
		None	Short	Long
Predicted	None	629	333	511
	Short	0	0	0
	Long	0	0	0

> What is the accuracy of our predictions in this confusion matrix?

$$\text{Accuracy} = \frac{629}{1473} = 0.427$$

(prevalence of the most common choice)

Class activity

> Do we do better than random guessing?

Sort of: we are really bad at predicting short-term use

$\Rightarrow$ probably want more explanatory variables in the model

Moral

- By itself, accuracy isn't particularly useful for summarizing prediction performance
- It is helpful to interpret accuracy in relation to simple random guessing. Our model isn't very good if we can't beat a random guess
- We also need to look at predictive ability for each class

Hypothesis testing

Research question: Is there a relationship between age and contraceptive choice?

> What are my steps to answer this question with a hypothesis test?

- specify a model for the relationship ✓
- specify hypotheses in terms of one or more model parameters (βs)
- calculate a test statistic
- calculate a p-value

Specify hypotheses

Research question: Is there a relationship between age and contraceptive choice?

$$\log\left(\frac{\pi_{i(Long)}}{\pi_{i(None)}}\right) = \beta_{0(Long)} + \beta_{1(Long)}Age_i + \beta_{2(Long)}Age_i^2$$

$$\log\left(\frac{\pi_{i(Short)}}{\pi_{i(None)}}\right) = \beta_{0(Short)} + \beta_{1(Short)}Age_i + \beta_{2(Short)}Age_i^2$$

> What should our null and alternative hypotheses be?

H_0: $\underbrace{\beta_{1(Long)} = \beta_{2(Long)} = \beta_{1(Short)} = \beta_{2(Short)}} = 0$

H_A: at least one is $\neq 0$

Specify hypotheses

Full model

$$\left\{\begin{array}{l}\log\left(\frac{\pi_{i(Long)}}{\pi_{i(None)}}\right) = \beta_{0(Long)} + \beta_{1(Long)}Age_i + \beta_{2(Long)}Age_i^2 \\ \log\left(\frac{\pi_{i(Short)}}{\pi_{i(None)}}\right) = \beta_{0(Short)} + \beta_{1(Short)}Age_i + \beta_{2(Short)}Age_i^2\end{array}\right.$$

$$H_0 : \beta_{1(Short)} = \beta_{2(Short)} = \beta_{1(Long)} = \beta_{2(Long)} = 0$$

$$H_A : \text{at least one of } \beta_{1(Short)}, \beta_{2(Short)}, \beta_{1(Long)}, \beta_{2(Long)} \neq 0$$

> What are the full and reduced models?

Reduced model:

$$\log\left(\frac{\pi_{i(Long)}}{\pi_{i(None)}}\right) = \beta_{0(Long)}$$

(intercept-only)

$$\log\left(\frac{\pi_{i(Short)}}{\pi_{i(None)}}\right) = \beta_{0(Short)}$$

Test statistic

> What test can I use to compare nested models?

LRT (testing multiple parameters)

G = deviance for reduced − deviance for full

$G \sim \chi^2_q$ q = # parameters tested

In our example, $q = 4$

Drop in deviance

```
m1 <- multinom(Choice ~ WifeAge + I(WifeAge^2),
               data = cmc_data)
```

```
summary(m1)
```

```
...
##
## Residual Deviance: 3015.821
...
```

Deviance for full model: 3015.821

> How would we fit the reduced model in R?

Choice ~ 1 ← fit an intercept-only model

Drop in deviance

```r
m0 <- multinom(Choice ~ 1,
               data = cmc_data)
```

```r
summary(m0)
```

```
...
##
## Residual Deviance: 3142.726
...
```

compare to χ^2_4

Reduced model deviance: 3142.726

Drop in deviance: $G = 3142.726 - 3015.821 = 126.905$

> What distribution do we use to calculate the p-value?

Calculating a p-value

Under $H_0, G \sim \chi^2_q$, where q is the number of parameters tested.

Here $q = 4$ (2 parameters for each log relative risk model)

```
pchisq(126.905, df=4, lower.tail=F)
```

```
## [1] 1.787184e-26
```

≈ 0

So we have very strong evidence that there is a relationship between age and contraceptive choice.