Lecture 26: Likelihood ratio tests

Asymptotics of the LRT

Generalization to higher dimensions

Earthquake data

Data from the 2015 Gorkha earthquake on 211774 buildings, with variables including:

- `Damage`: whether the building sustained any damage (1) or not (0)
- Age: the age of the building (in years)
- `Surface`: a categorical variable recording the surface condition of the land around the building. There are three different levels: n, o, and t

Likelihood ratio tests

```r
m1 <- glm(Damage ~ Age*Surface, data = earthquake,
          family = binomial)
summary(m1)$coefficients
```

```
                 Estimate  Std. Error      z value       Pr(>|z|)
(Intercept)   1.411099267 0.032512137  43.4022302  0.000000e+00
Age           0.059786157 0.002099615  28.4748245 2.401973e-178
Surfaceo      0.061461279 0.072860676   0.8435453  3.989236e-01
Surfacet     -0.474024473 0.034382357 -13.7868520  3.058165e-43
Age:Surfaceo  0.002807968 0.005087768   0.5519056  5.810130e-01
Age:Surfacet  0.008163407 0.002230082   3.6605868  2.516383e-04
```

We want to test whether the relationship between Age and Damage is the same for all three surface conditions. What hypotheses do we test?

Likelihood ratio tests

Full model:

```
m1 <- glm(Damage ~ Age*Surface, data = earthquake,
          family = binomial)
```

Reduced model:

```
m2 <- glm(Damage ~ Age + Surface, data = earthquake,
          family = binomial)
```

Comparing deviances

```
m1 <- glm(Damage ~ Age*Surface, data = earthquake,
          family = binomial)
m1$deviance
```

```
[1] 139150.5
```

```
m2 <- glm(Damage ~ Age + Surface, data = earthquake,
          family = binomial)
m2$deviance
```

```
[1] 139164.4
```

Comparing deviances

```
m1 <- glm(Damage ~ Age*Surface, data = earthquake,
          family = binomial)

m2 <- glm(Damage ~ Age + Surface, data = earthquake,
          family = binomial)

pchisq(m2$deviance - m1$deviance,
       m2$df.residual - m1$df.residual,
       lower.tail = F)
```

```
[1] 0.0009433954
```

Deviance

Definition: The *deviance* of a fitted model with parameter estimates $\widehat{\beta}$ is given by

$$2\ell(\text{saturated model}) - 2\ell(\widehat{\beta})$$

Summary: LRT for logistic regression