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

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

```
...
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.411099   0.032512  43.402  < 2e-16 ***
## Age           0.059786   0.002100  28.475  < 2e-16 ***
## Surfaceo      0.061461   0.072861   0.844 0.398924
## Surfacet     -0.474024   0.034382 -13.787  < 2e-16 ***
## Age:Surfaceo  0.002808   0.005088   0.552 0.581013
## Age:Surfacet  0.008163   0.002230   3.661 0.000252 ***
##
##     Null deviance: 153536  on 211773  degrees of freedom
## Residual deviance: 139150  on 211768  degrees of freedom
...
```

> 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)
```

Likelihood ratio tests

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

```
...
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.411099   0.032512  43.402  < 2e-16 ***
## Age             0.059786   0.002100  28.475  < 2e-16 ***
## Surfaceo        0.061461   0.072861   0.844 0.398924
## Surfacet       -0.474024   0.034382 -13.787  < 2e-16 ***
## Age:Surfaceo    0.002808   0.005088   0.552 0.581013
## Age:Surfacet    0.008163   0.002230   3.661 0.000252 ***
##
##     Null deviance: 153536  on 211773  degrees of freedom
## Residual deviance: 139150  on 211768  degrees of freedom
...
```

> What information replaces R^2 and R^2_{adj} in the GLM output?

Deviance

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

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

Residual and null deviance

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

```
...
##     Null deviance: 153536  on 211773  degrees of freedom
## Residual deviance: 139150  on 211768  degrees of freedom
...
```

Comparing deviances

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

```
...
##     Null deviance: 153536  on 211773  degrees of freedom
## Residual deviance: 139150  on 211768  degrees of freedom
...
```

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

```
...
##     Null deviance: 153536  on 211773  degrees of freedom
## Residual deviance: 139164  on 211770  degrees of freedom
...
```

> How should I use this output to calculate a test statistic?

Comparing deviances

```r
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
```

Summary: LRT for logistic regression