TwoSample t-test as a model comparison

André Meichtry

Contents

Simulation of data from model 1

Description 2

Classical test 2

Model approach 3
- Unrestricted model . . . 3
- Restricted ("Null") model . . . 3
- ANOVA for model comparison . . . 4
- By hand . . . 4
  - Residual sum of squares and explained sum of squares . . . 4
  - degrees of freedom . . . 4
  - F-test . . . 4
- Log-Likelihood of both models* . . . 5
- AIC and BIC (criterion for optimality)* . . . 5

Simulation of data from model

Simulate some two-group-data from model with parameters μ_i and σ assumed known:

- Means parameterization: $Y_{ij} = \mu_i + \epsilon_{ij}, i = 1, 2, j = 1, ..., n_i,$ $\boxed{Y_{ij} \sim N(\mu_i, \sigma^2)}$
- Effects parameterization (Default in R): $Y_{ij} = \alpha + I_{group=1}\beta + \epsilon_{ij}$ with $\mu_1 = \alpha$ and $\beta = \mu_2 - \mu_1$.

```
library(psych)
n <- 20
alpha <- 2
beta <- 3
sigma <-10
set.seed(10)
group <- as.factor(sample(c(0,1),n,replace=TRUE))
Y <- alpha+beta*(group==1)+rnorm(n,0,sigma)
headTail(data.frame(Y,group))
```

```
        Y group
1   13.02     0
2    9.56     0
3    2.62     1
4   14.87     1
...   ...  <NA>
17  -1.88     1
```

```
18   -3.72      1
19    3.98      1
20   -0.54      0
```

Description

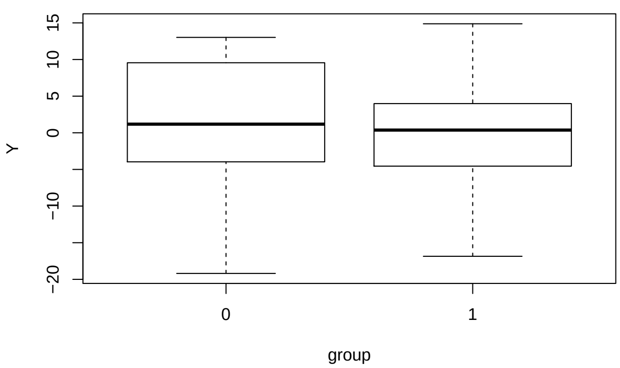
```r
summary(Y)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-19.191  -4.110   1.040   0.783   7.512  14.874 
```

```r
by(Y,group,summary)
```

```
group: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -19.19   -3.41    1.18    1.34    8.88   13.02 
------------------------------------------------------------ 
group: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-16.853  -4.343   0.371   0.229   3.749  14.874 
```

```r
boxplot(Y~group)
```



Classical test

```r
t.test(Y~group,var.equal=TRUE)
```

```

	Two Sample t-test

data:  Y by group
t = 0.261, df = 18, p-value = 0.8
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.8096 10.0280
sample estimates:
mean in group 0 mean in group 1 
        1.33773         0.22855 
```

Model approach

Unrestricted model

μ_1 and μ_2 are unknown and have to be estimated:

```
mod <- lm(Y~group)
summary(mod)
```

```
Call:
lm(formula = Y ~ group)

Residuals:
   Min     1Q Median     3Q    Max 
-20.53  -4.91  -0.16   6.17  14.65 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.34       3.00    0.45     0.66
group1         -1.11       4.25   -0.26     0.80

Residual standard error: 9.49 on 18 degrees of freedom
Multiple R-squared:  0.00378,   Adjusted R-squared:  -0.0516 
F-statistic: 0.0683 on 1 and 18 DF,  p-value: 0.797
```

```
confint(mod)
```

```
               2.5 % 97.5 %
(Intercept)  -4.9688 7.6443
group1      -10.0280 7.8096
```

Restricted (“Null”) model

$H_0 : \mu_2 - \mu_1 = 0$ (Means parametrization) or $H_0 : \beta = 0$ (Effects parameterization). In this case, we have only one mean to be estimated.

```
mod0 <- lm(Y~1)
summary(mod0)
```

```
Call:
lm(formula = Y ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.974  -4.893   0.257   6.729  14.091 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.783      2.070    0.38     0.71

Residual standard error: 9.26 on 19 degrees of freedom
```

ANOVA for model comparison

```
anova(mod0,mod)
```

```
Analysis of Variance Table

Model 1: Y ~ 1
Model 2: Y ~ group
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     19 1628                         
2     18 1622  1      6.15 0.07    0.8
```

By hand

Residual sum of squares and explained sum of squares

```
(RSS <-sum(mod$residuals^2))
```

```
[1] 1621.9
```

```
(RSS0 <-sum(mod0$residuals^2))
```

```
[1] 1628.1
```

```
(RSS0-RSS)
```

```
[1] 6.1514
```

Multiple R-squared. You find this quantity in the summary model output.

```
(RSS0-RSS)/RSS0
```

```
[1] 0.0037783
```

degrees of freedom

```
(df <- n-2)
```

```
[1] 18
```

```
(df0 <- n-1)
```

```
[1] 19
```

F-test

The F-statistic is the amount of available fit that is actually achieved,

$$F = \frac{(RSS_0 - RSS)/(df_0 - df)}{RSS/df} = \frac{"Explained\,MeanSquares"}{"Not\,explained\,MeanSquares"}$$

```
F <- (RSS0-RSS)/(df0-df)/(RSS/(df))
p <- 1-pf(F,df1=df0-df,df2=df)
sigma <- sqrt(RSS/df)
sigma0 <- sqrt(RSS0/df0)
print(data.frame(RSS0,RSS,SSExplained=RSS0-RSS,F,p,sigma,sigma0),row.names=FALSE)
```

```
   RSS0    RSS SSExplained        F       p  sigma sigma0
 1628.1 1621.9      6.1514 0.068268 0.79684 9.4925 9.2568
```

Log-Likelihood of both models*

These are the log-Likelhoods of the model at MLE's (the maximum likelihood estimates).

```r
logLik(mod)
```

```
'log Lik.' -72.335 (df=3)
```

```r
logLik(mod0)
```

```
'log Lik.' -72.373 (df=2)
```

AIC and BIC (criterion for optimality)*

Adding penalties for model complexity:

$$AIC = -2l + 2p$$

with l as the log-likelihood and p the number of parameters in the model.

$$BIC = -2l + 2\log(n)$$

```r
AIC(mod0,mod)
```

```
     df    AIC
mod0  2 148.75
mod   3 150.67
```

```r
BIC(mod0,mod)
```

```
     df    BIC
mod0  2 150.74
mod   3 153.66
```

Smaller AIC and BIC (smaller negative penalized likelihoood) are better. We do NOT reject the constrained model in favor of the unconstrained model.