NZSSN Courses: Introduction to R

Session 7 – Simple analysis

Statistical Consulting Centre

consulting@stat.auckland.ac.nz
The Department of Statistics
The University of Auckland

20 July, 2017

Regression commands

Two of the most commonly used `R` commands for modeling:

- `lm()`: fits **L**inear **M**odels
- `glm()`: fits **G**eneralised **L**inear **M**odels.\

Note SAS users: `PROC GLM` is **not** the same as `R`'s `glm()`.

There's a lot in these two commands; entire stage 3 statistical courses on linear and generalised linear models.

Student's t-test

`t.test(y ~ x)`

- `y`: values; e.g., `Cholesterol`, `BMI`, `Age`, etc.
- `x`: group; e.g., `Sex`, `Smoke.group`.

Suppose we want to test whether males and females (`x = Sex`) have different Cholesterol levels.

Categorical variables should be converted to type `factor` before analysis, i.e.

```
combined.long.df$Sex <- factor(combined.long.df$Sex)
with(combined.long.df, t.test(Cholesterol ~ Sex))
```

Student's t-test

```
##
##  Welch Two Sample t-test
##
## data:  Cholesterol by Sex
## t = 11.029, df = 48066, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equ
## 95 percent confidence interval:
##  3.723005 5.332270
## sample estimates:
## mean in group Female   mean in group Male
##             208.1640             203.6364
```

- p-value $< 2.2e\text{-}16$.
- We have extremely strong evidence that the cholesterol level for male is different from female.

Multiple comparisons

Let's compare the total score between three age groups, i.e.

1. Do a t-test between "Under 35" and "36 to 60".
2. Do a t-test between "Under 35" and "Over 61".
3. Do a t-test between "36 to 60" and "Over 61".

Really?

Error rate

When we do a t-test comparing mean total score between females and males, the null hypothesis is that the mean total score for females is the same as that for males. The t-test is performed (with the hope) to reject this null hypothesis.

In order to come up with a p-value, we *assume* that α (typically 5%) of the time, we will reject the null hypothesis when it's actually true, i.e., we assume 5% of the time we will make a mistake.

- When we do two simultaneous t-tests, about 10% of the time we will make a mistake.
- When we do three simultaneous t-tests, about 15% of the time we will make a mistake.
- The chance of being shot in Russian Roulette is 16.67%. Would you risk it then?

Analysis ***o***f ***Va***riance (ANOVA)

Generalises t-test to more than two groups

Null hypothesis: all group means are equal.

Example. Mean Cholesterol level is the same for all three age.groups.

```
tryaov <- with(combined.long.df, aov(Cholesterol~Age.group))
```

- `aov()`: **A**nalysis **o**f **V**ariance.
- Response variable (i.e. `total.lik`) is separated by ~ from explanatory variable(s) (i.e. `age.group`).
- All explanatory variables should be categorical (otherwise it's not ANOVA).

aov()

```
summary(tryaov)
```

```
##                 Df   Sum Sq Mean Sq F value Pr(>F)    
## Age.group        2 10038041 5019020    2731 <2e-16 ***
## Residuals   48183 88566029    1838                    
## ---
## Signif. codes:  
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2904 observations deleted due to missingness
```

We have extremely strong evidence that at least one age group's mean Cholesterol level is different to that of the other age groups.\

Which one(s) is(are) different????

Which one(s)?

```
model.tables(tryaov, "means")
```

```
## Tables of means
## Grand mean
##
## 206.0412
##
##  Age.group
##     Under 35 36 to 60 Over 61
##        186.3    210.7   221.2
## rep  15780.0  17040.0 15366.0
```

The mean Cholesterol level...

- over all participants is 206.

Which one(s)?

```
model.tables(tryaov, "means")
```

```
## Tables of means
## Grand mean
##
## 206.0412
##
##  Age.group
##     Under 35 36 to 60 Over 61
##        186.3    210.7   221.2
## rep  15780.0  17040.0 15366.0
```

The mean Cholesterol level...

- for "Under 35" group is lower than both that of the "36 to 60" and the "Over 61" groups.
- for "36 to 60" group is lower than the "Over 61" group.

Which one(s)?

```r
model.tables(tryaov, "means")
```

```
## Tables of means
## Grand mean
##
## 206.0412
##
##  Age.group
##     Under 35 36 to 60 Over 61
##        186.3    210.7   221.2
## rep  15780.0  17040.0 15366.0
```

Are any pairs of these means statistically different from one another?

Post-hoc multiple comparisons

```r
TukeyHSD(tryaov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Cholesterol ~ Age.group)
##
## $Age.group
##                           diff       lwr      upr p adj
## 36 to 60-Under 35 24.37127 23.261145 25.48138     0
## Over 61-Under 35  34.88304 33.744215 36.02186     0
## Over 61-36 to 60  10.51177  9.393915 11.62963     0
```

- `diff`: estimated difference between two group means.
- `lwr, upr`: lower and upper limit of the 95% confidence interval of the estimated difference.
- `p adj`: p-values adjusted for multiple comparisons.

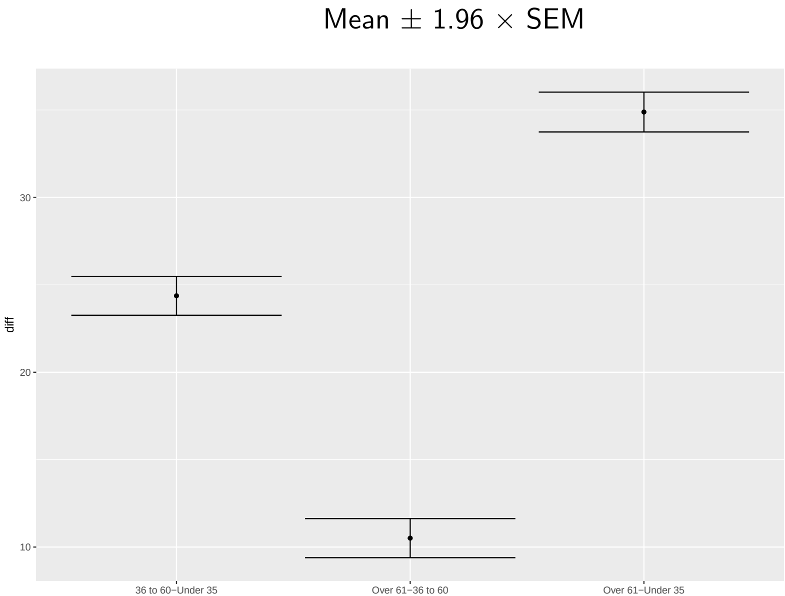
Post-hoc multiple comparisons

```r
comp <- TukeyHSD(tryaov)
comp$Age.group
```

```
##                          diff       lwr      upr p adj
## 36 to 60-Under 35 24.37127 23.261145 25.48138     0
## Over 61-Under 35  34.88304 33.744215 36.02186     0
## Over 61-36 to 60  10.51177  9.393915 11.62963     0
```

- Mean Cholesterol level for "36 to 60" is 24.4 mg/100ml *higher* than "Under 35" (`p adj` < 0.0001).
- Mean Cholesterol level for "Over 61" is 34.9 mg/100ml *higher* than "Under 35" (`p adj` < 0.0001).
- Mean Cholesterol level for "Over 61" is 10.5 mg/100ml *higher* than "36 to 60" (`p adj` < 0.0001).

From Session 6: Mean Cholesterol level vs Age group



Two-way ANOVA

- `tryaov` was fitted using one categorical explanatory variable (`Age.group`). We therefore refer to its ANOVA table as *one-way*.
- If we fit a linear model using two categorical explanatory variables, we have a *two-way* ANOVA.
- Recall: All categorical variables should be converted into factors.

```
combined.long.df$Sex <- factor(combined.long.df$Sex)
try2way <- with(combined.long.df,
                aov(Cholesterol~Sex*Age.group))
```

- `Sex*Age.group` is equivalent to `Sex + Age.group + Sex:Age.group`.

Two-way ANOVA

```r
summary(try2way)
```

```
##                   Df   Sum Sq Mean Sq F value Pr(>F)
## Sex                1   245990  245990   136.6 <2e-16 ***
## Age.group          2 10076421 5038210  2797.8 <2e-16 ***
## Sex:Age.group      2  1519391  759696   421.9 <2e-16 ***
## Residuals      48180 86762267    1801
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2904 observations deleted due to missingness
```

There is two-way interaction between `Sex` and `Age.group` (p-value $= 0.19$), i.e., the magnitude of the difference in mean Cholesterol levels between males and females is not constant across all age groups, and vice versa.

Estimated means

```r
model.tables(try2way, "means")
```

```
## Tables of means
## Grand mean
##
## 206.0412
##
##  Sex
##      Female    Male
##       208.2   203.6
## rep 25593.0 22593.0
##
##  Age.group
##     Under 35 36 to 60 Over 61
##        186.3    210.6   221.2
## rep  15780.0  17040.0 15366.0
##
```

Estimated means

```r
model.tables(try2way, "means")$table$'Sex:Age.group'
```

```
##          Age.group
## Sex       Under 35 36 to 60 Over 61
##   Female 184.7069 209.6103 231.4095
##   Male   188.1431 211.8985 210.1405
```

Post-hoc pairwise comparisons

```r
TukeyHSD(try2way)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Cholesterol ~ Sex * Age.group)
##
## $Sex
##                  diff       lwr       upr p adj
## Male-Female -4.527637 -5.286902 -3.768373     0
##
## $Age.group
##                       diff       lwr      upr p adj
## 36 to 60-Under 35 24.37080 23.272004 25.46959     0
## Over 61-Under 35  34.96254 33.835337 36.08974     0
## Over 61-36 to 60  10.59174  9.485293 11.69819     0
##
```

Post-hoc pairwise comparisons

```r
TukeyHSD(try2way)$`Sex:Age.group`
```

```
##                                        diff       lwr
## Male:Under 35-Female:Under 35     3.4361409  1.505591
## Female:36 to 60-Female:Under 35  24.9033236 23.079721
## Male:36 to 60-Female:Under 35    27.1915865 25.299622
## Female:Over 61-Female:Under 35   46.7026213 44.815819
## Male:Over 61-Female:Under 35     25.4335479 23.508474
## Female:36 to 60-Male:Under 35    21.4671826 19.570071
## Male:36 to 60-Male:Under 35      23.7554456 21.792531
## Female:Over 61-Male:Under 35     43.2664804 41.308541
## Male:Over 61-Male:Under 35       21.9974070 20.002561
## Male:36 to 60-Female:36 to 60     2.2882629  0.430431
## Female:Over 61-Female:36 to 60   21.7992977 19.946724
## Male:Over 61-Female:36 to 60      0.5302244 -1.361314
## Female:Over 61-Male:36 to 60     19.5110348 17.591130
## Male:Over 61-Male:36 to 60       -1.7580386 -3.715568
```

Test of independence

```
smoke.age.tab <- with(combined.df, table(Smoke.group, Age.grou
smoke.age.tab
```

```
##              Age.group
## Smoke.group Under 35 36 to 60 Over 61
##          No      643     1548    2064
##         Yes     1732     1840     799
```

Do smoking habit depend on age group? Statistically speaking, is `Smoke.group` and `Age.group` independent of one another?

Pearson's Chi-squared test

```
chisq.test(smoke.age.tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  smoke.age.tab
## X-squared = 1082.1, df = 2, p-value < 2.2e-16
```

- There is extremely strong evidence (p-value < 0.0001) that `Smoke.group` and `Age.group` are not independent of one another.
- Smoking habit depend on the age group to which patient belong.

Assumptions

- Pearson's Chi-squared tests have certain assumptions. Beyond the scope of this course. \item `chisq.test()` will give you a warning if these assumptions are not met.

```
## Warning in chisq.test(mytest): Chi-squared approximation ma
## be incorrect

##
##  Chi-squared test for given probabilities
##
## data:  mytest
## X-squared = 2, df = 3, p-value = 0.5724
```

- These assumptions are more likely to be wrong if the sample size is small.
- If this happens, the alternative is to use Fisher's exact test.

Fisher's exact test

Assume `Q5.age.tab` does not meet the underlying assumptions of Pearson's Chi-squared test.

```r
fisher.test(smoke.age.tab, simulate.p.value = TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated
##  p-value (based on 2000 replicates)
##
## data:  smoke.age.tab
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

Summary

- Student's t-test
- One-way ANOVA
- Two-way ANOVA
- Pearson's Chi-squared test
- Fisher's exact test