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

2 March, 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., `total.lik`, `Q1.lik`, `Age`, etc.
- `x`: group; e.g., `Gender`, `Q5` (`obedient` or `think themselves`).

Suppose we want to test whether males and females (`x = Gender`) have different total scores across `Q1` – `Q4` (`y = total.lik`).

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

```
issp.df$Gender <- factor(issp.df$Gender)
with(issp.df, t.test(total.lik~Gender))
```

Student's *t*-test

```
Welch Two Sample t-test

data:  total.lik by Gender
t = 4.3417, df = 874.71, p-value = 1.579e-05
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 0.3541459 0.9384793
sample estimates:
mean in group Female   mean in group Male
            12.71067             12.06436
```

- The estimated difference in total score between females and males is $12.71 - 12.06 = 0.65$.
- p-value $= 1.579 \times 10^{-5}$, i.e. we have extremely strong evidence that the mean total score are statistically significantly different.

Student's *t*-test

```
Welch Two Sample t-test

data:  total.lik by Gender
t = 4.3417, df = 874.71, p-value = 1.579e-05
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 0.3541459 0.9384793
sample estimates:
mean in group Female   mean in group Male
            12.71067             12.06436
```

- While we have statistical significance, we should note that the sample sizes are very large.
- Is the observed difference significant from a social scientist's perspective?

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 of Variance (ANOVA)

Generalises *t*-test to more than two groups.

Null hypothesis: all group means are equal.

Example. Mean total score is the same for all three age.groups.

```
tryaov <- with(issp.df, aov(total.lik~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    803   401.4   89.82 <2e-16 ***
Residuals   951   4250     4.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
93 observations deleted due to missingness
```

We have extremely strong evidence that at least one age group's mean total score 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

12.43711

 age.group
    Under 35 36 to 60 Over 61
       13.39    12.46   10.79
rep   319.00   446.00  189.00
```

The mean total score...

- over all participants is 12.4.
- for “Under 35” group is higher than both that of the “36 to 60” and the “Over 61” groups.
- for ”36 to 60” group is higher than the ”Over 61” group.

Which one(s)?

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

Tables of means
Grand mean

12.43711

 age.group
    Under 35 36 to 60 Over 61
       13.39    12.46   10.79
rep   319.00   446.00  189.00
```

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

Post-hoc multiple comparisons

```
TukeyHSD(tryaov)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = total.lik ~ age.group)

$age.group
                              diff       lwr        upr p adj
36 to 60-Under 35 -0.9335578 -1.297433 -0.5696823     0
Over 61-Under 35  -2.6003549 -3.055857 -2.1448527     0
Over 61-36 to 60  -1.6667972 -2.097496 -1.2360986     0
```

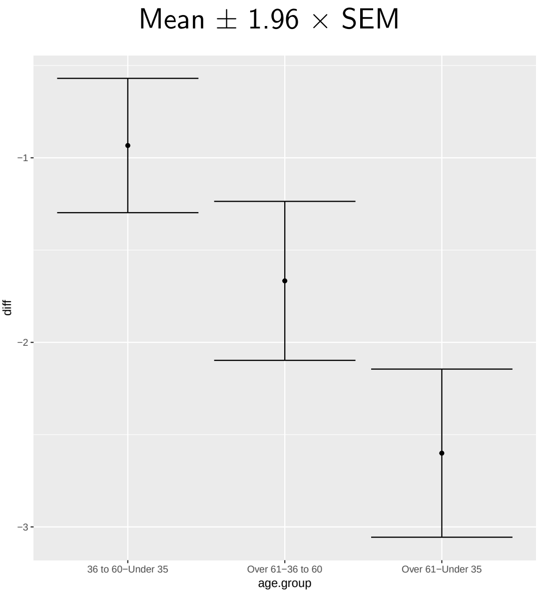
- `diff`: estimated difference between two group means.
- `lwr`, `upr`: lower and upper limit of the 95% confidence interval of the estimated difference.
- `p adj`: p-values adjusted for multiple comparisons.

Post-hoc multiple comparisons

```
                         diff       lwr        upr        p adj
36 to 60-Under 35 -0.9335578 -1.297433 -0.5696823 7.353107e-09
Over 61-Under 35  -2.6003549 -3.055857 -2.1448527 1.399991e-13
Over 61-36 to 60  -1.6667972 -2.097496 -1.2360986 1.690870e-13
```

- Mean total score for "36 to 60" is 0.9 units (on the likert scale) *lower* than "Under 35" (`p adj` < 0.0001).
- Mean total score for "Over 61" is 2.6 units *lower* than "Under 35" (`p adj` < 0.0001).
- Mean total score for "Over 61" is 1.7 units *lower* than "36 to 60" (`p adj` < 0.0001).

From Session 6: Mean total score 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.

```
issp.df$Gender <- factor(issp.df$Gender)
try2way <- with(issp.df,
                aov(total.lik~Gender*age.group))
```

- `Gender*age.group` is equivalent to `Gender + age.group + Gender:age.group`.

Two-way ANOVA

```
summary(try2way)

                  Df Sum Sq Mean Sq F value   Pr(>F)    
Gender             1     98    97.7  22.200 2.82e-06 ***
age.group          2    774   386.8  87.905  < 2e-16 ***
Gender:age.group   2     15     7.3   1.654    0.192    
Residuals        947   4167     4.4                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
94 observations deleted due to missingness
```

There is no two-way interaction between `Gender` and `age.group` (p-value $= 0.19$), i.e., the magnitude of the difference in mean total score between males and females is constant across all age groups, and vice versa.

Two-way ANOVA

```
summary(try2way)

                 Df Sum Sq Mean Sq F value   Pr(>F)    
Gender            1     98    97.7  22.200 2.82e-06 ***
age.group         2    774   386.8  87.905  < 2e-16 ***
Gender:age.group  2     15     7.3   1.654    0.192    
Residuals       947   4167     4.4                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
94 observations deleted due to missingness
```

We have extremly strong evidence that:

- the mean total score of *at least one* age group differs from the others, and
- mean total score differs between males and females.

Estimated means

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

Tables of means
Grand mean

12.43757

 Gender
    Female   Male
     12.71  12.06
rep 549.00 404.00

 age.group
    Under 35 36 to 60 Over 61
       13.39    12.43   10.84
rep   319.00   445.00  189.00
```

Post-hoc pairwise comparisons

```
TukeyHSD(try2way)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = total.lik ~ Gender * age.group)

$Gender
                  diff        lwr        upr   p adj
Male-Female -0.6478476 -0.9176817 -0.3780135 2.8e-06

$age.group
                        diff       lwr        upr p adj
36 to 60-Under 35 -0.9533867 -1.314615 -0.5921584     0
Over 61-Under 35  -2.5488847 -3.000862 -2.0969079     0
Over 61-36 to 60  -1.5954980 -2.023006 -1.1679900     0
```

Test of independence

```
Q5.age.tab <- with(issp.df, table(Q5, age.group))
Q5.age.tab

                  age.group
Q5                  Under 35 36 to 60 Over 61
  be obedient             38       74      75
  think themselves       259      353     122
```

Do opinions on preparing children for life depend on age group?
Statistically speaking, is `Q5` (the variable) and `age.group` independent of one another?

Pearson's Chi-squared test

```
chisq.test(Q5.age.tab)



Pearson's Chi-squared test


data:  Q5.age.tab
X-squared = 51.115, df = 2, p-value = 7.955e-12
```

- There is extremely strong evidence (p-value < 0.0001) that `Q5` and `age.group` are not independent of one another.
- Opinions on preparing children for life depend on the age group to which respondents belong.

Assumptions

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

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

- 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.

```
fisher.test(Q5.age.tab)



Fisher's Exact Test for Count Data


data:  Q5.age.tab
p-value = 6.93e-11
alternative hypothesis: two.sided
```

Summary

- Student's t-test
- One-way ANOVA
- Two-way ANOVA
- Pearsons Chi-squared test
- Fishers exact test