Summary of Homework

周睿

2022-12-14

目录


```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
```

```
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##
## Attaching package: 'rstatix'

## The following object is masked from 'package:stats':
##
##     filter

##
## bruceR (version 0.8.9)
## BRoadly Useful Convenient and Efficient R functions
##
## Packages also loaded:
## √ dplyr      √ emmeans       √ ggplot2
## √ tidyr      √ effectsize    √ ggtext
## √ stringr    √ performance   √ cowplot
## √ forcats    √ lmerTest      √ see
## √ data.table
##
## Main functions of `bruceR`:
## cc()             Describe()  TTEST()
## add()            Freq()      MANOVA()
## .mean()          Corr()      EMMEANS()
## set.wd()         Alpha()     PROCESS()
## import()         EFA()       model_summary()
## print_table()    CFA()       lavaan_summary()
##
## https://psychbruce.github.io/bruceR/
##
```

```
## These R packages are dependencies of `bruceR` but not installed:
## pacman, lmtest, vars, phia, GGally, GPArotation
## ***** Please Install All Dependencies *****
## install.packages("bruceR", dep=TRUE)
```

1 t test

1.1 判断题

1.1.1 方差同质性与检验量

- 独立样本 t 检验中，假设两组样本数据的容量 n 相同，并且数据均值差异 μ_D 不变。如果两组数据的方差不同质，和方差同质的情况相比，更容易检出显著差异。
- 判断：错误
- 原因：方差同质的情况下，检验的误差更小，更容易检出差异。并且如果方差不同质，校正自由度后更不容易显著。
- 有一组不满足方差齐性的数据，某同学对这组数据进行独立样本 t 检验时，忘记了做方差同质性检验而误认为数据满足方差齐性，那么本次假设检验犯第一类错误的概率会增大。
- 判断：正确 ## 计算题 ### 方差同质性检验 ### 独立样本检验的不同类型 ### 独立样本检验与配对样本检验的比较 # one-way ANOVA ## 判断题 ### ANOVA 与 t 检验的联系
- 如果使用单因素方差分析比较 K(K>2) 组数据能够拒绝 H0，那么将这 K 组数据两两进行独立样本 t 检验，一定存在某两组数据的均值存在显著差异。
- 判断：正确
- 原因：这里是不矫正的条件。ANOVA 相当于是对 t 检验的显著性水平的校正，不校正时，t 检验存在显著差异是 ANOVA 存在显著差异

的必要不充分条件。

- 在同样的显著性水平下，对两组数据分别进行独立样本 t 检验与两个水平的单因素方差分析 (one-way anova)，若 t 检验的结果显著，那么方差分析的结果也显著。
- 判断：正确
- 原因：one-way anova 依赖于两个样本方差同质的假设，而独立样本 t 检验在方差同质和方差不同质的情况下都可以开展。在方差同质的前提下，one-way anova 和独立样本 t 检验能得到相同的结果；但是在方差不同质的假设下，若 t 检验的结果显著，方差同质的检验也应该显著，故命题正确。

1.1.2 分组变量

当进行方差分析时，只有独立因子（independ factor）才能作为分组变量，半独立因子（quasi-independent factor）不可以作为分组变量。

- 判断：错误
- 原因：独立因子和半独立因子都可以作为分组变量半独立因子举例：性别、地区、季节等

1.1.3 学习效应

如果关心的研究问题是 “不同复习策略对心理统计课程期末考试成绩的影响”，则最好使用单因素方差分析而非重复测量方差分析来处理数据。

- 判断：正确
- 原因：存在练习效应不适合用重复测量

1.1.4 重复测量更易显著

重复测量方差分析比单因素方差分析更容易检验出显著性差异的可能原因之一是：组数和每组人数都一样的条件下，重复测量方差分析拒绝 H0 的临

界值更小

- 判断：错误
- 原因：重复测量 F 统计量的分母自由度小，临界值更大。重复测量更容易显著的原因是，这能够剔除个体差异，让检验更有效力，如果 treatment effect 显著，更容易被检验出。

2 two-way ANOVA

2.1 判断题

2.1.1 交互作用理解

一个 2 × 3 的二因素独立测量方差分析的主效应有 2 个，交互作用有 6 个。（注：2 × 3 表示两个因素分别有 2 个、3 个水平）

- 判断：错误
- 原因：交互作用只有一个，即为 A*B
- 练习：2x3x3 设计的交互作用有 4 个

2.1.2 主效应和简单主效应理解

在二因素独立测量方差分析中，交互作用显著时需检验简单主效应，即不考虑其中一个因素，只检验另一因素不同水平的两两之间的差异。

- 判断：错误
- 原因：检验简单主效应是在，分别固定其中一个因素在它的各个水平上，检验另一因素处于不同水平的两两之间的差异。题目中描述的是主效应的事后比较。

2.1.3 事后比较的显著性水平校正

在二因素独立测量方差分析中，其中一个因素有 5 个水平，若对该水平进行简单主效应检验，总体显著性水平 =0.05，使用 Bonferroni 方法校正显著性，则无论研究假设是什么，每次比较的调整后显著性水平都是 0.05 ÷ 10 = 0.005。

- 判断：错误
- 原因：结合研究假设，事后捉对比较的组数可能并不等于穷举的数量，也即可能并不用检验 10 次，使用 `bonferroni` 方法校正时，可能" 分母" 并不是 10.（比如选取其中一组进行参照，那么只要比较四次）

2.2 计算题

```
load('2W_ANOVA.Rdata')
```

2.2.1 分组变量处理

某研究者采用两次单因素方差分析探究了以下两个问题，请完成两次方差分析**并使用规范的文字报告结果**（8 分）。

- 问题一：不同开放性（分为高、中、低 3 个水平）被试实验得分的不同。
- 问题二：被试性别与旁观者性别的差异性（被试与旁观者性别相同、不同时）被试实验得分的不同。

```
mydata <- score %>%
     select(O,gender,audience, result) %>%
     mutate(O_lvl = cut(O,c(0.5,2.5,4.5,6.5),labels = c('low','median','high')),
            gender_diff = as.factor(ifelse(gender == audience, 'Same','Diff'))) %>%
  select(O_lvl,gender_diff,result)
```

3 mixed-design ANOVA

3.1 判断题

3.1.1 模式理解

假设我们要进行 2x3 的方差分析，因素 A 是组间变量有 A1，A2，A3 三个水平，因素 B 是组内变量有 B1，B2 三个水平，在每个实验情境（condition）下有 10 名被试。要完成上述实验设计，应该招募 30 名被试。

注意：组间变量的数据都由一个被试生成！因此招募被试要看组间因素的水平组合！

- 判断：正确
- 理由：每组人数 * 组间条件数 =30

3.1.2 F 检验量的理解

如果错误地把 B 看成组间变量，那么对因素 B 的主效应进行检验选择的 F 分布临界值会比真实值小

- 判断：**正确**
- 理由：把 B 当成组间变量时，df_{error} 会比把 B 当成组内变量时的更大，因为后者剔除了个体差异，对应的 F 检验量的第二个自由度会变得更小，在相同的显著性水平下，临界值会大。

3.1.3 事后检验与因素水平

接上。如果 A 和 B 的交互作用显著，控制 A 后对 B 的简单主效应进行检验，需要进行 3 次 F 检验；如果简单主效应显著，需要进行简单主效应的事后检验。

- 判断：错误
- 理由：A 有三个水平，检验简单主效应时是要进行三次 F 检验，但由于 B 因素只有两个水平，因此如果显著，没必要进行事后检验。

3.2 计算题

```
# ID 表示被试编号
# Collectivism 表示被试在集体主义-个人主义取向上的得分，高分偏向集体主义，低分偏向个人主义
# Level 是组间变量，表示视觉比较任务的难度，Hard 代表复杂任务，Easy 代表简单任务
# Accuracy_01 表示被试独立进行视觉比较判断的正确率（基线水平）
# Accuracy_04,Accuracy_08,Accuracy_16 分别表示总参与者（包含被试自己）人数为 4，8，16 时，
load('confirmity.RData')
```

3.2.1 结果汇报

- 请选择合适的方法进行方差同质性的检验和校正
- 统计方法：说明变量个数、变量类型（组内/组间）、每个变量的水平数、方差分析名称（Twoway/混合设计等）
- 主效应和交互作用报告：(F, p, η^2), 结论是：
- 事后检验：如果需要，报告矫正方法，差异显著的组间报告统计量或 p 值，差异不显著的组可以不报告
- 简单主效应：如果需要，报告在 A1/A2/A3 水平下，对变量 B 主效应的检验，和（3）格式一致，如果简单主效应显著，根据需要判断是否需要进一步事后检验；
- 简单用易读的语言概括结论，回应题目的问题。言之有理、格式规范即可得分

报告样例（以本题为例）：

实验设计 2x4 混合设计方差分析：任务难度是组间变量，有简单和困难两个水平；参与人数是组内变量，有 1，4，8，16 四个水平。

Anova 的结果表明（两个主效应和一个交互作用）：

- 任务难度对正确率影响的主效应显著 $F(1, 38) = 174.074, p < 0.001, \eta^2 = 0.821$, 不需要事后检验；
- 参与者人数对正确率影响的主效应不显著 $F(3, 114) = 0.374, p = 0.772, \eta^2 = 0.010$, 不需要事后检验；

- 任务难度和参与者人数对正确率影响的交互作用显著 $F(3, 114) = 7.210, p < 0.001, \eta^2 = 0.159$, 需要简单主效应检验

控制任务难度的具体水平进行简单主效应检验的结果表明（思考：为什么这里不做控制参与者人数比较任务难度的简单主效应）：

- 简单任务中，参与者人数对正确率影响的主效应显著 $F(3, 38) = 2.996, p = 0.043, \eta^2 = 0.191$, 需要进行事后检验；
- 困难任务中，参与者人数对正确率影响的主效应显著 $F(3, 38) = 3.549, p = 0.023, \eta^2 = 0.219$, 需要进行事后检验；

使用 sidak 校正的事后检验比较简单任务中不同参与者人数对正确率的影响：

- 简单任务：与基线水平相比，8 人 (p=0.067) 条件和 16 人条件 (p=0.089) 下正确率与基线水平的差异达到了**边缘显著**，其他参与条件与基线水平相比不存在显著差异
- 复杂任务：4 人条件正确率高于基线水平 (p=0.0195)

综合以上：困难任务中没有出现从众效应，而简单任务中出现了典型的从众效应，参与者任务大于 1 时的判断正确率显著低于基线水平。

3.2.2 数据转化与统计

请将 collectivism 得分转化为分组变量，大于中位数的编码为 “集体主义”，小于等于中位数的编码为 “个人主义”，报告简单任务组中 “集体主义” 和 “个人主义” 被试的人数（10 分）。

```
# 数据转化 method1
confirmity_grouped = confirmity
confirmity_grouped$group = ifelse(
  confirmity$collectism > median(confirmity$collectism),
  'collectism',
  'individualism'
)
```

```r
# 数据转化 method2
confirmity_grouped = mutate(
  confirmity,
  group = ifelse(collectism > median(collectism), 'collectism', 'individualism')
)

# 计数 method1
confirmity_grouped %>%
  filter(level == 'easy') %>%
  group_by(group) %>%
  count()

# 计数 method2
easy_group = confirmity[confirmity_grouped$group == 'collectism' & confirmity_grouped$l
hard_group = confirmity[confirmity_grouped$group == 'collectism' & confirmity_grouped$l

nrow(easy_group)
nrow(hard_group)
```

3.2.3 混合设计方差分析

```r
# 处理宽数据
model = bruceR::MANOVA(
  data = confirmity,
  subID = 'ID',
  dvs = c('accuracy_01', 'accuracy_04', 'accuracy_08', 'accuracy_16'),
  dvs.pattern = 'accuracy_(..)',
  between = 'level',
  within = 'participant'
  ## within 参数相当于是对组内变量赋予名字，仅在宽数据如此
)
```

```
##
## ====== ANOVA (Mixed Design) ======
##
## Descriptives:
##
##  "level" "participant"   Mean    S.D.  n
##
##     easy participant01 93.250 (0.967) 20
##     easy participant04 90.600 (0.883) 20
##     easy participant08 90.250 (1.118) 20
##     easy participant16 90.700 (2.658) 20
##     hard participant01 81.350 (4.705) 20
##     hard participant04 84.850 (4.626) 20
##     hard participant08 83.650 (4.146) 20
##     hard participant16 83.850 (4.671) 20
##
## Total sample size: N = 40
##
## ANOVA Table:
## Dependent variable(s):      accuracy_01, accuracy_04, accuracy_08, accuracy_16
## Between-subjects factor(s): level
## Within-subjects factor(s):  participant
## Covariate(s):               -
##
##                             MS    MSE df1 df2       F     p      ²p [90% CI of  ²p]
##
## level                2418.025 13.891   1  38 174.074 <.001 ***   .821 [.731, .872] .
## participant             4.042 10.797   3 114   0.374  .772       .010 [.000, .032] .
## level * participant    77.842 10.797   3 114   7.210 <.001 ***   .159 [.057, .251] .
##
## MSE = mean square error (the residual variance of the linear model)
##  ²p = partial eta-squared = SS / (SS + SSE) = F * df1 / (F * df1 + df2)
##  ²p = partial omega-squared = (F - 1) * df1 / (F * df1 + df2 + 1)
```

```
##  ²G = generalized eta-squared (see Olejnik & Algina, 2003)
## Cohen' s f² =  ²p / (1 -  ²p)
##
## Levene' s Test for Homogeneity of Variance:
##
##                   Levene' s F df1 df2     p
##
## DV: accuracy_01       27.392   1  38 <.001 ***
## DV: accuracy_04       26.408   1  38 <.001 ***
## DV: accuracy_08       19.638   1  38 <.001 ***
## DV: accuracy_16        3.553   1  38  .067 .
##
##
## Mauchly' s Test of Sphericity:
##
##                       Mauchly's W     p
##
## participant                0.9348  .780
## level * participant        0.9348  .780
##
```

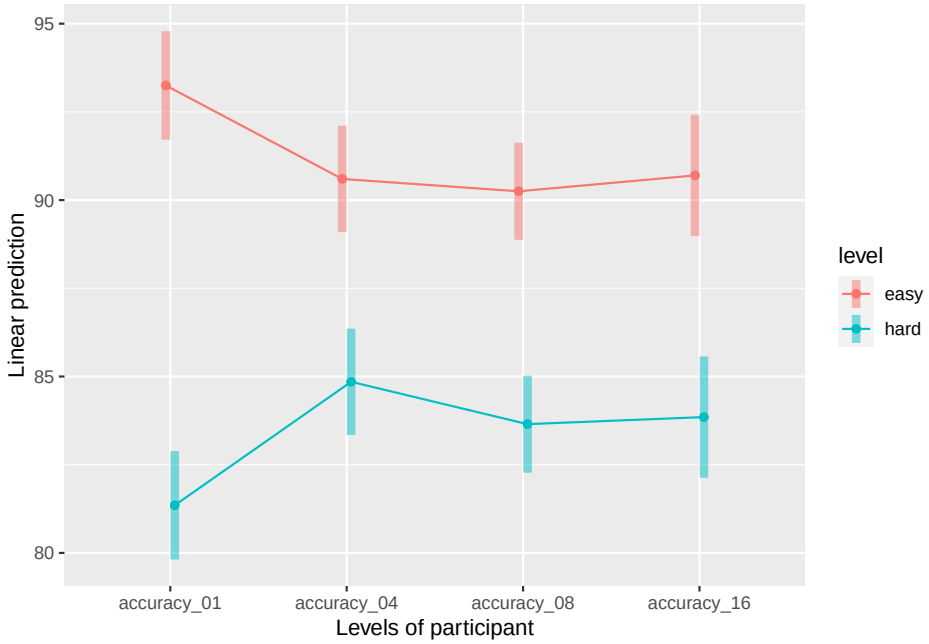
```
# 处理长数据
## 指定不转换的列为 id，剩下的宽转长
long_data = reshape2::melt(confirmity, id = list('ID', 'level', 'collectism'))
## 指定何为变量，何为值
long_data = rename(long_data, participant = variable, accuracy = value)
model = bruceR::MANOVA(
  long_data,
  subID = 'ID',
  dv = 'accuracy',
  between = 'level',
  within = 'participant'
)
```

```
##
##     * Data are aggregated to mean (across items/trials)
##     if there are >=2 observations per subject and cell.
##     You may use Linear Mixed Model to analyze the data,
##     e.g., with subjects and items as level-2 clusters.

##
## ====== ANOVA (Mixed Design) ======
##
## Descriptives:
##
##  "level" "participant"   Mean    S.D.  n
##
##     easy   accuracy_01 93.250 (0.967) 20
##     easy   accuracy_04 90.600 (0.883) 20
##     easy   accuracy_08 90.250 (1.118) 20
##     easy   accuracy_16 90.700 (2.658) 20
##     hard   accuracy_01 81.350 (4.705) 20
##     hard   accuracy_04 84.850 (4.626) 20
##     hard   accuracy_08 83.650 (4.146) 20
##     hard   accuracy_16 83.850 (4.671) 20
##
## Total sample size: N = 40
##
## ANOVA Table:
## Dependent variable(s):      accuracy
## Between-subjects factor(s): level
## Within-subjects factor(s):  participant
## Covariate(s):               -
##
##                             MS    MSE df1 df2       F     p      ²p [90% CI of  ²p]
##
## level                2418.025 13.891   1  38 174.074 <.001 ***   .821 [.731, .872] .
## participant             4.042 10.797   3 114   0.374  .772       .010 [.000, .032] .
```

```
## level * participant    77.842 10.797   3 114   7.210 <.001 ***   .159 [.057, .251] .
##
## MSE = mean square error (the residual variance of the linear model)
##  ²p = partial eta-squared = SS / (SS + SSE) = F * df1 / (F * df1 + df2)
##  ²p = partial omega-squared = (F - 1) * df1 / (F * df1 + df2 + 1)
##  ²G = generalized eta-squared (see Olejnik & Algina, 2003)
## Cohen' s f² =  ²p / (1 -  ²p)
##
## Levene' s Test for Homogeneity of Variance:
##
##                  Levene' s F df1 df2     p
##
## DV: accuracy_01      27.392   1  38 <.001 ***
## DV: accuracy_04      26.408   1  38 <.001 ***
## DV: accuracy_08      19.638   1  38 <.001 ***
## DV: accuracy_16       3.553   1  38  .067 .
##
##
## Mauchly' s Test of Sphericity:
##
##                      Mauchly's W     p
##
## participant               0.9348  .780
## level * participant       0.9348  .780
##
```

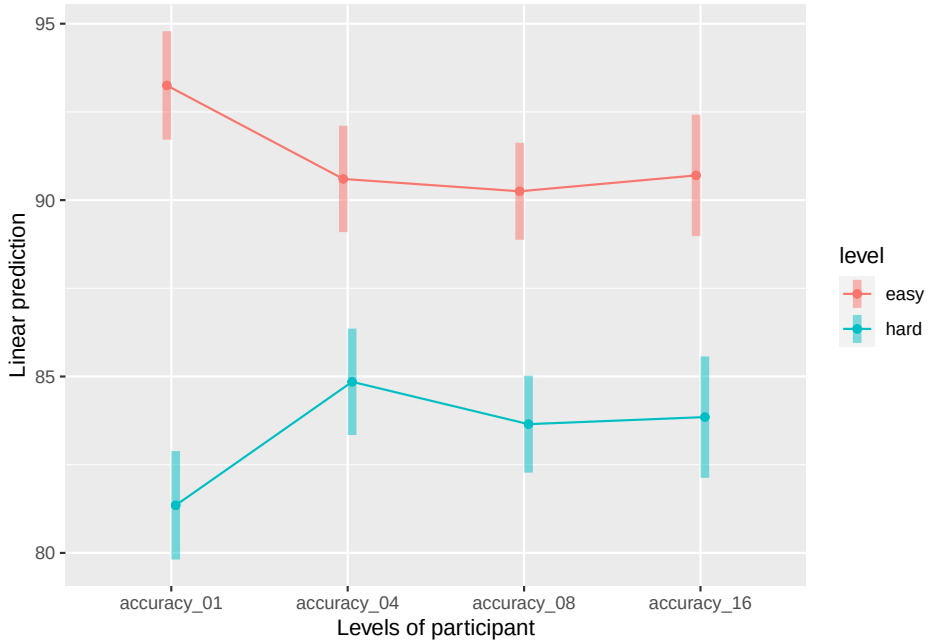
```
emmeans::emmip(model, level ~ participant, CIs = TRUE)
```



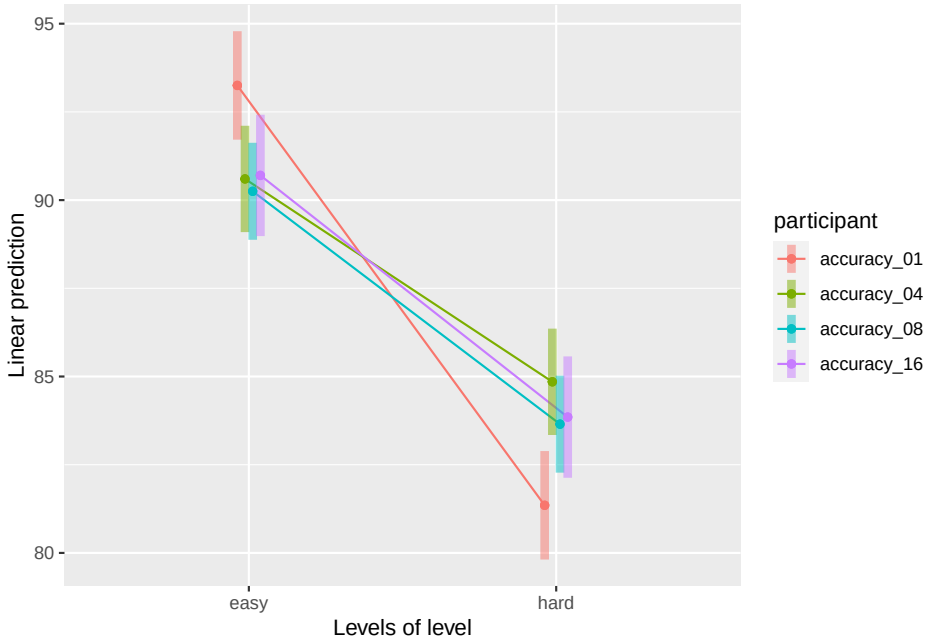
3.2.4 两个因素作用图示

`emmip(model,formula,CIs)`

```
emmeans::emmip(
  model,
  level ~ participant,
  CIs = TRUE
)
```




```
emmeans::emmip(
  model,
  participant~level,
  CIs = TRUE
)
```



3.2.5 数据筛选与连续方差分析

```
confirmity %>%
  filter(level == 'easy') %>%
  MANOVA(
       subID = 'ID',
       dvs = c('accuracy_01','accuracy_04','accuracy_08','accuracy_16'),
       dvs.pattern = 'accuracy_(..)',
       within = 'participant') %>%
  EMMEANS(effect = 'participant')
```

```
##
## ====== ANOVA (Within-Subjects Design) ======
##
## Descriptives:
##
##  "participant"   Mean    S.D.  n
```

```
##
##  participant01 93.250 (0.967) 20
##  participant04 90.600 (0.883) 20
##  participant08 90.250 (1.118) 20
##  participant16 90.700 (2.658) 20
##
## Total sample size: N = 20
##
## ANOVA Table:
## Dependent variable(s):      accuracy_01, accuracy_04, accuracy_08, accuracy_16
## Between-subjects factor(s): -
## Within-subjects factor(s):  participant
## Covariate(s):               -
##
##                  MS   MSE df1 df2      F     p      ²p [90% CI of  ²p]   ²G
##
## participant  38.100 2.302   3  57 16.553 <.001 ***   .466 [.294, .580] .375
##
## MSE = mean square error (the residual variance of the linear model)
##  ²p = partial eta-squared = SS / (SS + SSE) = F * df1 / (F * df1 + df2)
##  ²p = partial omega-squared = (F - 1) * df1 / (F * df1 + df2 + 1)
##  ²G = generalized eta-squared (see Olejnik & Algina, 2003)
## Cohen' s f² =  ²p / (1 -  ²p)
##
## Levene' s Test for Homogeneity of Variance:
## No between-subjects factors. No need to do the Levene' s test.
##
## Mauchly' s Test of Sphericity:
##
##              Mauchly's W     p
##
## participant       0.1557 <.001 ***
##
```

```
## The sphericity assumption is violated.
## You may specify: sph.correction="GG" (or ="HF")
##
## ------ EMMEANS (effect = "participant") ------
##
## Joint Tests of "participant":
##
##       Effect df1 df2      F     p      ²p [90% CI of  ²p]
##
##  participant   3  19 61.354 <.001 ***   .906 [.824, .941]
##
## Note. Simple effects of repeated measures with 3 or more levels
## are different from the results obtained with SPSS MANOVA syntax.
##
## Estimated Marginal Means of "participant":
##
##  "participant"   Mean [95% CI of Mean]    S.E.
##
##  participant01 93.250 [92.798, 93.702] (0.216)
##  participant04 90.600 [90.187, 91.013] (0.197)
##  participant08 90.250 [89.727, 90.773] (0.250)
##  participant16 90.700 [89.456, 91.944] (0.594)
##
##
## Pairwise Comparisons of "participant":
##
##                       Contrast Estimate    S.E. df       t     p     Cohen' s d [95%
##
##  participant04 - participant01   -2.650 (0.254) 19 -10.426 <.001 *** -1.235 [-1.584,
##  participant08 - participant01   -3.000 (0.281) 19 -10.677 <.001 *** -1.398 [-1.784,
##  participant08 - participant04   -0.350 (0.335) 19  -1.046 1.000     -0.163 [-0.622,
##  participant16 - participant01   -2.550 (0.609) 19  -4.187  .003 **  -1.188 [-2.024,
##  participant16 - participant04    0.100 (0.710) 19   0.141 1.000      0.047 [-0.928,
```

```
##  participant16 - participant08    0.450 (0.500) 19   0.900 1.000      0.210 [-0.476,
##
## Pooled SD for computing Cohen' s d: 2.146
## P-value adjustment: Bonferroni method for 6 tests.
##
## Disclaimer:
## By default, pooled SD is Root Mean Square Error (RMSE).
## There is much disagreement on how to compute Cohen' s d.
## You are completely responsible for setting `sd.pooled`.
## You might also use `effectsize::t_to_d()` to compute d.
```

4 相关系数

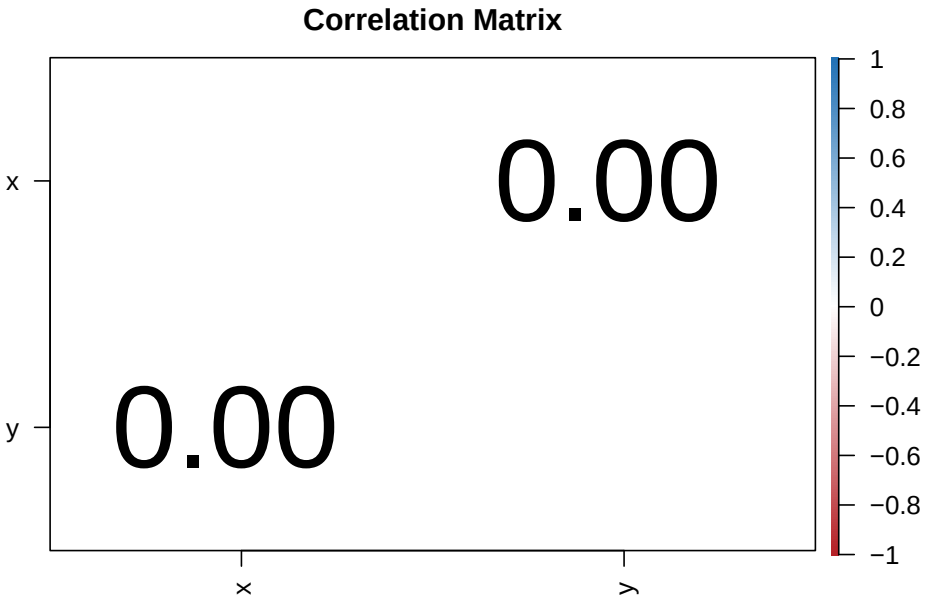
4.1 判断题

4.1.1 线性相关

如果两个变量线性相关系数为皮尔逊相关系数 r 为 0，那么可以认为两个变量没有相关关系。

- 判断：错误
- 原因：皮尔逊相关系数 r 刻画的是线性相关关系，皮尔逊相关系数 r 为 0 只能代表没有线性相关关系，而不能说明没有相关关系。比如，变量 y 关于变量 x 是二次函数关系，r=0。

```
x=seq(-2,2,length.out=100)
y=x^2
Corr(data.frame(x,y))
```



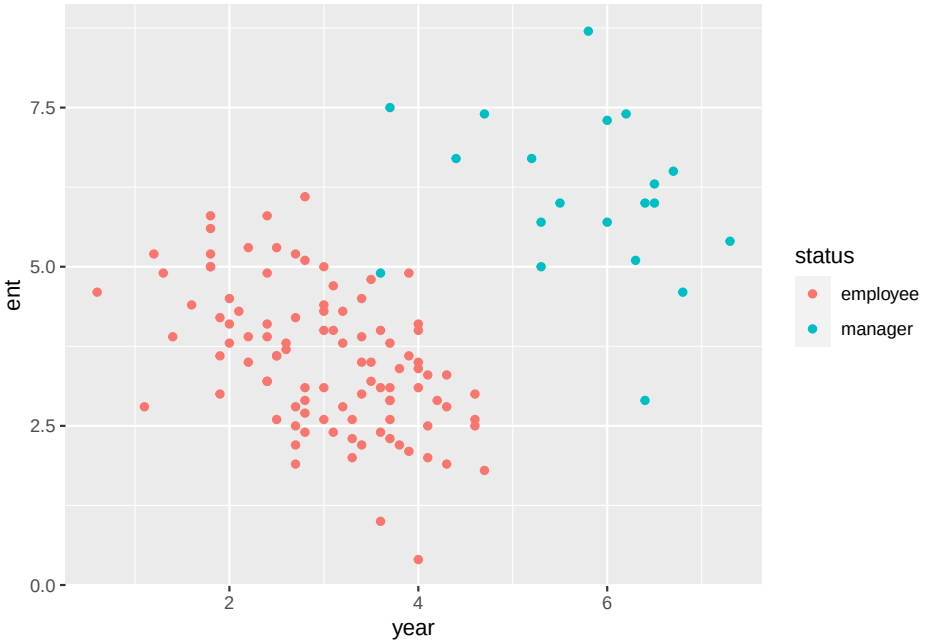

```
## Correlation matrix is displayed in the RStudio `Plots` Pane.
##
## Pearson's r and 95% confidence intervals:
##
##         r      [95% CI]     p       N
##
## x-y  0.00 [-0.20, 0.20] 1.000     100
##
```

4.2 计算题

4.2.1 数据分类与相关系数计算

使用 `filter` 函数

```
load('HW11.Rdata')
enthusiam %>%
  ggplot(aes(year,ent,color = status)) +
  geom_point()
```




```
enthusiam %>%
  filter(status == 'manager') %>%
  cor.test(~year+ent,data=.)
```

```
##
##  Pearson's product-moment correlation
##
## data:  year and ent
## t = -1.0748, df = 18, p-value = 0.2966
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6206543  0.2209489
## sample estimates:
##        cor
## -0.2455837
```

```
enthusiam %>%
  filter(status == 'employee') %>%
  cor.test(~year+ent,data=.)
```

```
##
##  Pearson's product-moment correlation
##
## data:  year and ent
## t = -5.8732, df = 98, p-value = 5.866e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6422871 -0.3487753
## sample estimates:
##        cor
## -0.5102391
```

5 线性回归

5.1 判断题

5.1.1 决定系数

对于 X,Y 两列变量的回归方程，决定系数 (R-squared) 表示 Y 的变异性中可以被 XY 的关系解释的部分的比例。

- 判断：正确
- 解释：例如，假如 r^2=0.6，说明总变异中有 0.6 的变异可以用 X,Y 的差异来解释

5.1.2 非线性相关

时刻注意数据之间可能存在的非线性关系!

使用最小二乘法 (OLS) 拟合 X,Y 两列变量得到线性回归方程，对回归方程的显著性进行 F 检验的结果是不显著，则 X,Y 不存在相关关系。

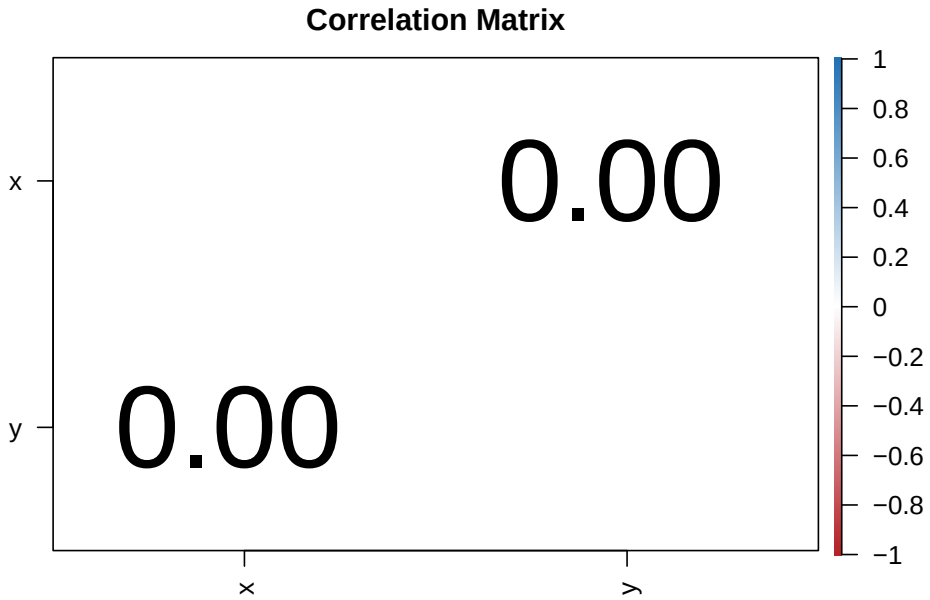
- 判断：错误

- 原因：最小二乘法拟合得到的线性回归方程，尝试解释并预测的是线性相关关系。如果结果不显著，只能说明 X,Y 无线性相关关系（严谨地说，是在给定的显著性水平下不能拒绝"X,Y 不存在相关关系"），但不能说明 X,Y 没有其他相关关系。比如二次函数关系，F 检验不显著，p 值为 1，但 X,Y 存在确定的相关关系。

```
x=seq(-1,1,0.05)
y=x^2
summary(lm(y~x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3500 -0.2875 -0.1000  0.2125  0.6500
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.500e-01  5.008e-02   6.988 2.22e-08 ***
## x           1.905e-16  8.466e-02   0.000        1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3207 on 39 degrees of freedom
## Multiple R-squared:  6.991e-32,  Adjusted R-squared:  -0.02564
## F-statistic: 2.727e-30 on 1 and 39 DF,  p-value: 1
```

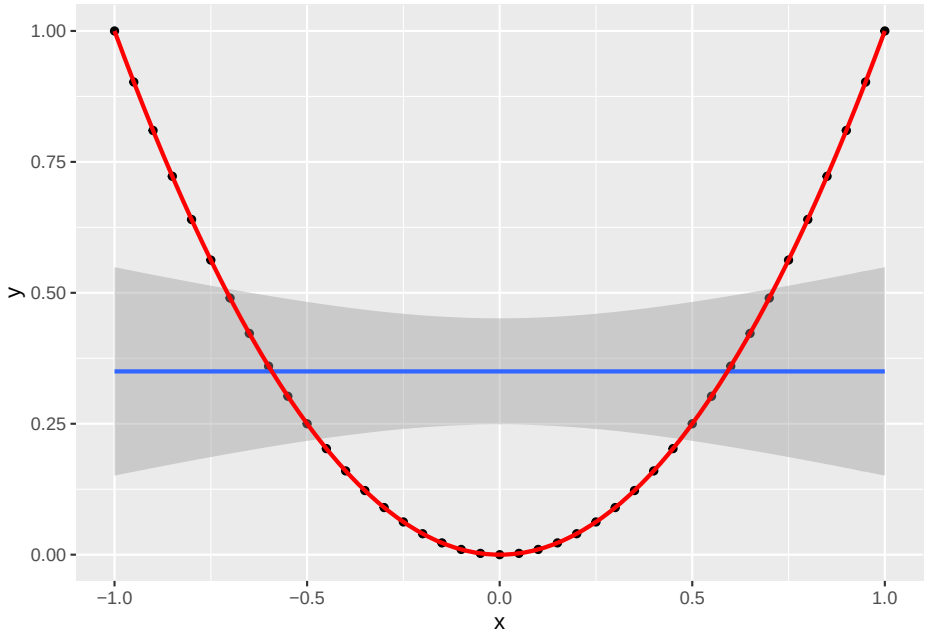
```
bruceR::Corr(data.frame(x,y))
```




```
## Correlation matrix is displayed in the RStudio `Plots` Pane.
##
## Pearson's r and 95% confidence intervals:
##
##         r      [95% CI]     p      N
##
## x-y  0.00 [-0.31, 0.31] 1.000     41
##
```

```
ggplot(data = data.frame(x,y),mapping = aes(x=x,y=y))+
  geom_point()+
  geom_smooth(method = 'lm')+
  geom_smooth(color='red')
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



5.1.3 模型检验与相关检验的关系

使用最小二乘法拟合 X,Y 两列变量得到线性回归方程，对回归方程的显著性进行 F 检验，检验结果的 p 值与两列变量 Pearson 相关的显著性检验的 p 值一致。

- 判断：正确

5.1.4 线性回归的数据类型要求

一元线性回归的数据要求因变量为等距或比率数据，在实际操作中，如果有足够多的水平，顺序数据也可以作为因变量。

- 判断：正确

5.2 计算题

```
# 有研究者欲研究大学生的睡眠时间与 GPA 的关系，从某高校选取了部分同学的数据
load('hw12.Rdata')
```

5.2.1 结果报告

5.2.1.1 相关系数 一定要说明相关系数的类型 (spearman,pearson) 和大小，同时报告假设检验的 t（包含自由度）和 p 值

5.2.1.2 线性模型 回归模型的方程 + 显著性检验的报告（需要包含 F、p 值、决定系数、调整后的决定系数）

5.2.2 异常值的筛选与清除

```
# 根据 3 规则，确定索引遮罩（由 T&F 构成）
x_mean = mean(corr$sleep)
x_sd = sd(corr$sleep)
outliers_mask = !(abs(corr$sleep - x_mean) > x_sd * 3)
## ! 表示反选

# 利用遮罩进行索引即可
corr[outliers_mask,]
```

5.2.2.1 使用 3 规则

```
##        sleep  gpa
## 1   5.792934 1.60
## 2   7.277429 1.37
## 3   8.084441 3.44
## 5   7.429125 3.21
## 6   7.506056 2.00
```

```
## 7   6.425260 1.37
## 8   6.453368 0.76
## 9   6.435548 1.37
## 10  6.109962 1.93
## 11  6.522807 2.54
## 12  6.001614 1.94
## 13  6.223746 0.97
## 14  7.064459 2.36
## 15  7.959494 0.83
## 16  6.889715 1.61
## 17  6.488990 1.60
## 18  6.088805 0.41
## 19  6.162828 2.43
## 20  9.415835 1.70
## 21  7.134088 1.66
## 22  6.509314 2.80
## 23  6.559452 2.28
## 24  7.459589 1.86
## 25  6.306280 2.16
## 26  5.551795 1.98
## 27  7.574756 3.11
## 28  5.976344 1.95
## 29  6.984862 2.24
## 30  6.064051 2.01
## 31  8.102298 1.74
## 32  6.524407 1.90
## 33  6.290560 2.95
## 34  6.498742 1.21
## 36  5.832381 2.69
## 38  5.659007 0.98
## 39  6.705706 1.24
## 40  6.534102 1.55
## 41  8.449496 1.45
```

```
## 42  5.931357 1.57
## 43  6.144635 1.49
## 44  6.719377 1.72
## 45  6.005660 2.05
## 46  6.031486 2.20
## 47  5.892682 2.92
## 48  5.748014 1.68
## 49  6.476172 1.60
## 50  6.503150 2.85
## 52  6.417924 1.84
## 53  5.891110 1.51
## 54  5.985038 0.48
## 55  6.837690 2.85
## 56  7.563056 1.36
## 57  8.647817 1.28
## 58  6.226647 4.00
## 59  8.605910 2.22
## 60  5.842191 1.72
## 61  7.656588 0.04
## 62  9.548991 2.08
## 63  6.965240 2.56
## 64  6.330366 2.09
## 65  6.992395 2.52
## 66  8.777084 3.46
## 67  5.861392 2.56
## 68  8.367827 1.68
## 69  8.329565 2.52
## 70  7.336473 1.78
## 71  7.006893 1.51
## 72  6.544531 0.00
## 73  6.633476 1.29
## 74  7.648287 2.30
## 75  9.070271 3.62
```

```
## 76  6.846602 2.22
## 77  5.609299 2.15
## 78  6.276418 1.12
## 79  7.258262 2.03
## 80  6.682941 0.39
## 81  6.822210 2.20
## 82  6.830006 2.35
## 83  5.627698 1.84
## 84  6.826213 2.35
## 85  7.850232 2.20
## 86  7.697609 2.49
## 87  7.549997 3.24
## 88  6.597268 2.61
## 89  6.808406 1.88
## 90  5.805472 0.96
## 91  6.946841 2.17
## 92  7.255196 1.63
## 93  8.705964 3.11
## 94  8.001513 0.87
## 95  6.504417 1.53
## 96  7.355550 2.22
## 97  5.865392 1.64
## 98  7.878204 1.56
## 99  7.972917 1.78
## 100 9.121117 2.55
## 101 7.414524 2.56
## 102 6.525282 2.29
## 103 7.065993 3.48
## 104 6.497522 2.65
## 105 6.174001 1.99
## 106 7.166989 1.44
## 107 6.103735 3.82
## 108 7.168185 2.38
```

```
## 109 7.354968 1.44
## 110 6.947895 2.01
## 111 6.804065 1.63
## 112 6.350930 1.61
## 113 5.890233 1.09
## 114 7.849274 2.06
## 115 7.022363 2.11
## 116 7.831141 1.22
## 117 5.755712 0.00
## 118 7.169026 1.72
## 119 7.673166 1.98
## 120 6.973724 2.30
```

```
# 采用 identify_outliers 函数发现异常值
outliers=corr %>% identify_outliers(gpa)
# 剔除原始数据中的异常值
corr_normal=corr[!corr[,c('sleep','gpa')] %in% outliers[,c('sleep','gpa')]]
```

5.2.2.2 使用 identify_outliers() 函数

5.2.3 基本手算

- 方差拆分与各项计算
- 决定系数的计算（包含调整后决定系数）
- 对模型的 F 检验