Reference Solutions

A Preprint

Taojun Hu *
Department of Biostatistics
Peking University

2011110158@stu.pku.edu.cn

November 9, 2021

Abstract

1 Homework

- chapter 2: page 244, 3(2); page 246 4; page 247 3(1) and page 248 4(1)
- chapter 3: page 249 3(1)(2)(4)
- chapter 4: page 254 3(3)(4); page 256 3(1)(4)
- chapter 5: page 258 1; page 258 2, 3

2 Reference solutions

```
# please first library the following packages: tidyverse, ggpubr
# if (! require(pacman)) install.packages("pacman")
# pacman::p_load(tidyverse, ggpubr)
```

2.1 Solutions for chapter 2

1. (Page 244 3(2))

```
get.per <- function(lower, upper){
  pnorm(upper, mean = 146, sd = 8) - pnorm(lower, mean = 146, sd = 8)}
get.per(138, 154)
```

```
## [1] 0.6826895
```

```
get.per(130, 162)
```

```
## [1] 0.9544997
```

2. (Page 246 4)

*If you find any errors including typos, it's welcomed to contact me by email

```
data.co <- data.frame(transfer = c('no', 'yes'), a_case = c(45, 710),
                      a_survive = c(35, 450), b_case = c(300, 83),
                      b_survive = c(215, 42)) %>%
    mutate(a_rate = a_survive/a_case, b_rate = b_survive/b_case) %>%
    mutate(total = a_case + b_case) %>%
    mutate(a_surv = a_rate*total, b_surv = b_rate*total)

data.simplify <- data.co %>% dplyr::select(c('total', 'a_surv', 'b_surv')) %>%
    colSums()
a_std_surv = data.simplify[2]/data.simplify[1]
b_std_surv = data.simplify[3]/data.simplify[1]
cat('Standard survival rate for hospital A: ', a_std_surv, '\n')
```

```
## Standard survival rate for hospital A:  0.6774508
```

```
cat('Standard survival rate for hospital B: ', b_std_surv, '\n')
```

```
## Standard survival rate for hospital B:  0.5698832
```

According to standard survival rate, we can't assert that hospital B has a higher survival rate than hospital A.

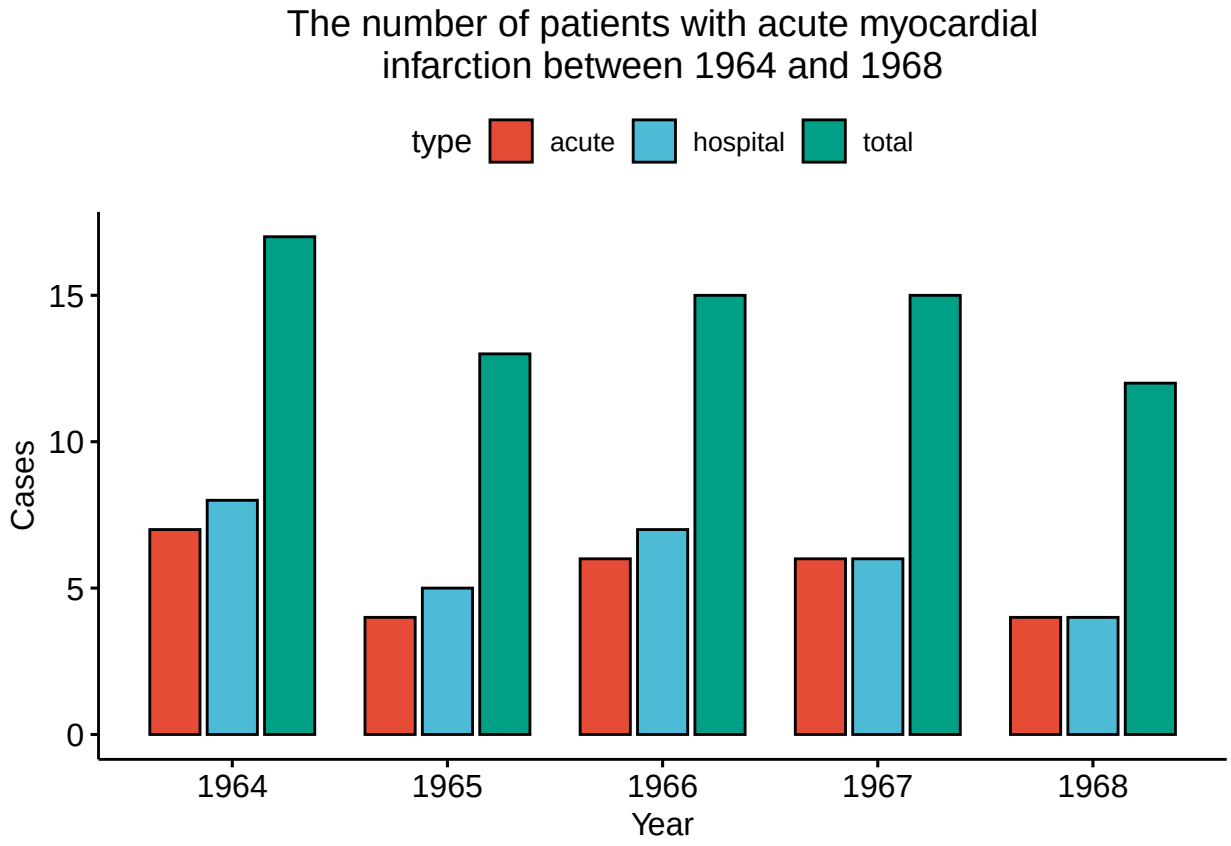
3. (Page 247 3(1))

Omitted.

4. (Page 248 4(1))

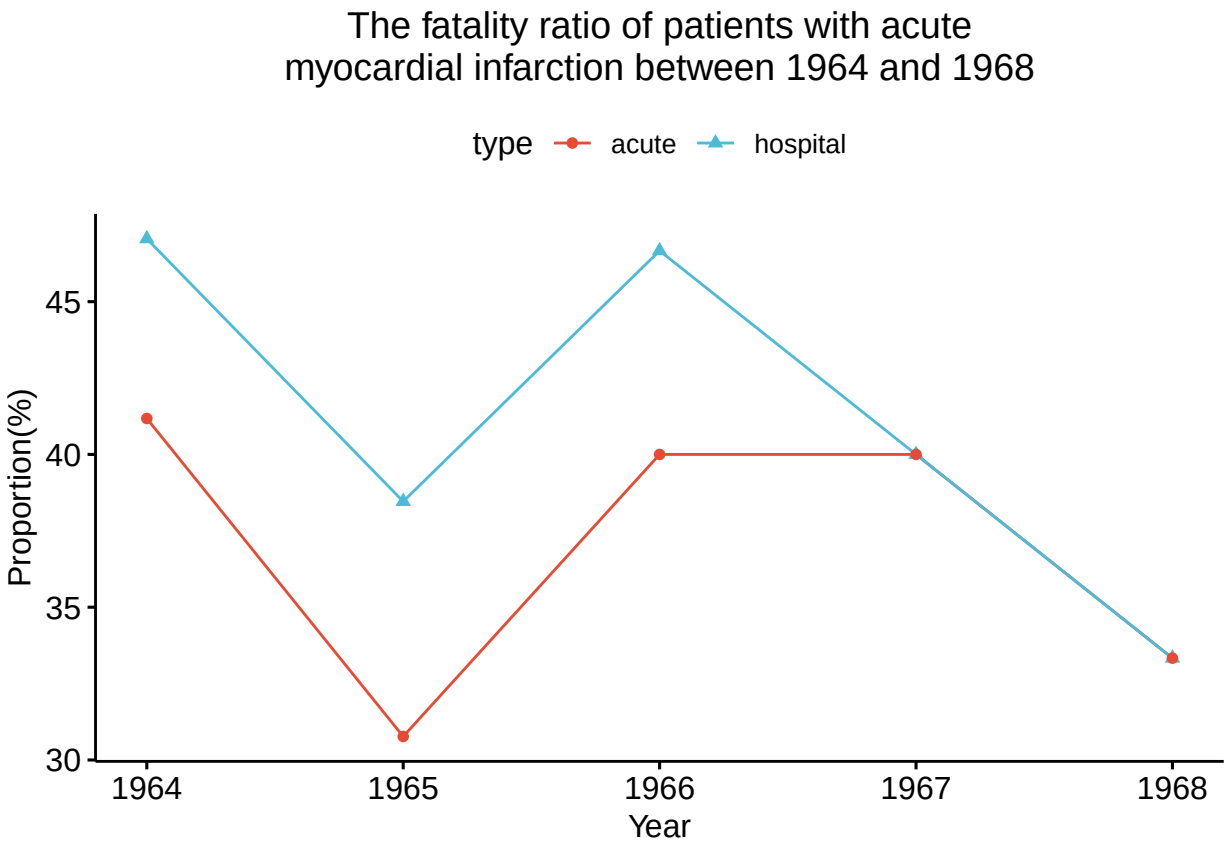
```
data.number <- data.frame(year = 1964:1968, total = c(17, 13, 15, 15, 12),
                          hospital = c(8, 5, 7, 6, 4), acute = c(7, 4, 6, 6, 4))
data.number.tidy <- data.number %>%
    pivot_longer(cols = !year, names_to = "type", values_to = "count") %>%
    mutate(type = factor(type))

ggbarplot(data = data.number.tidy, x = 'year', y = 'count', fill = 'type', merge = TRUE,
          palette = 'npg', xlab = 'Year', ylab = 'Cases',
          title = "The number of patients with acute myocardial
          infarction between 1964 and 1968" %>%
            stringr::str_wrap(width = 50)) + theme(plot.title = element_text(hjust = .5))
```




```
data.proportion <- data.number %>%
  mutate(hospital = hospital/total*100, acute = acute/total*100) %>%
  dplyr::select(!total) %>%
  pivot_longer(cols = !year, names_to = 'type', values_to = 'count')

ggline(data = data.proportion, x = 'year', y = 'count', shape = 'type', color = 'type',
       palette = 'npg', xlab = 'Year', ylab = 'Proportion(%)',
       title = "The fatality ratio of patients with acute myocardial
       infarction between 1964 and 1968" %>% stringr::str_wrap(width = 50)) +
  theme(plot.title = element_text(hjust = .5))
```



2.2 Solutions for chapter 3

1. (Page 249 3(1))

```
conf.d <- function(mu, sd, n) qt(.975, df = n-1)*c(-1, 1)*sd/sqrt(n) + mu
cat('Confidence interval for 1st sample\n')
```

```
## Confidence interval for 1st sample
```

```
conf.d(6.39, 2.24, 20)
```

```
## [1] 5.341648 7.438352
```

```
cat('confidence interval for 2nd sample\n')
```

```
## confidence interval for 2nd sample
```

```
conf.d(6.45, 2.51, 93)
```

```
## [1] 5.933072 6.966928
```

Sample 2 has a shorter confidence interval compared with sample 1. Sample 1 is more reliable because of larger sample size, shorter confidence interval.

2. (Page 249 3(2))

```
t.stat <- sqrt(360)*(4.66 - 4.84)/0.58
pt(t.stat, df = 360-1, lower.tail = TRUE)
```

```
## [1] 4.482972e-09
```

The p-value is less than 0.05, suggesting that there is significant difference.

Caution: this is a one-side hypothesis !

3. (Page 249 3(4))

```
barx <- c(11.6, 6.9)
sdx <- c(7.3, 2.7)
varx <- sdx**2
n <- 40
cat('p-value for equal variance testing\n')
```

```
## p-value for equal variance testing
```

```
pf(varx[1]/varx[2], df1 = n-1, df2 = n-1, lower.tail = FALSE)
```

```
## [1] 4.333099e-09
```

We can't reckon that two samples have identical variance. We use Satterthwaite approximation to test mean level, where under H_0, $T = \frac{\bar{X}-\bar{Y}}{\sqrt{S_1^2/n_1+S_2^2/n_2}} \approx t(m^*)$ with $m^* = (S_1^2/n_1 + S_2^2/n_2)^2/[\frac{1}{n_1-1}(S_1^2/n_1)^2 + \frac{1}{n_2-1}(S_2^2/n_2)^2]$.

```
m <- ( (sum(varx)/n)**2/(1/(n - 1)*sum((varx/n)**2)) )%>%round()
t.stat <- (barx[1] - barx[2])/sqrt(sum(varx/n))
cat('p-value is\n')
```

```
## p-value is
```

```
2*pt(t.stat, df = m, lower.tail = FALSE)
```

```
## [1] 0.0003773703
```

There is significant difference between two groups as p-value less than 0.05.

2.3 Solutions for chapter 4

1. (Page 254, 3(3))

```
p.0 <- 0.2
n.0 <- 400
x_lower <- qnorm(0.95)*sqrt(n.0*p.0*(1 - p.0)) + n.0*p.0
x_lower %>% ceiling()
```

```
## [1] 94
```

The above code uses normal approximation, while the following code calculates exact p-value

```
qbinom(0.95, 400, prob = 0.2) %>% ceiling()
```

```
## [1] 93
```

Caution: it is a one-side hypothesis testing with $H_0 : p > 0.2$

2. (Page 254, 3(1))

```
poisson.test(225, r = 100*2, alternative = 'greater')
```

```
##
##  Exact Poisson test
##
## data:  225 time base: 1
## number of events = 225, time base = 1, p-value = 0.04361
## alternative hypothesis: true event rate is greater than 200
## 95 percent confidence interval:
##  200.9087      Inf
## sample estimates:
## event rate
##        225
```

With a p-value less than 0.05, we reject the null hypothesis and conclude that the water is *unqualified*.

Caution: sum of two independent poisson distribution does also follow a poisson distribution; it's also a one-side hypothesis testing

3. (Page 256, 3(1))

```
data.drug <- matrix(c(28, 18, 10, 9, 20, 24), nrow = 3)
chisq.test(data.drug, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data.drug
## X-squared = 15.556, df = 2, p-value = 0.0004189
```

Significant difference.

4. (Page 256, 3(4))

```
data.co <- matrix(c(120*0.35, 120*(0.6 - 0.35),
                    120*(0.5 - 0.35), 120*(1 - 0.35 - 0.25 - 0.15)),
                  nrow = 2)
mcnemar.test(data.co)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  data.co
## McNemar's chi-squared = 2.5208, df = 1, p-value = 0.1124
```

No significant differences.

2.4 Solutions for chapter 5

1. (Page 258, 1)

```
data.thyroid <- data.frame(group = c(rep('high', 9), rep('mid', 8), rep('low', 7)) %>% factor(),
                           thyroid = c(34, 45, 49, 55, 58, 59, 60, 72, 86, 8, 25, 36,
                                       40, 42, 53, 65, 74, 5, 8, 18, 32, 45, 47, 65))

bartlett.test(thyroid ~ group, data = data.thyroid)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  thyroid by group
## Bartlett's K-squared = 1.1653, df = 2, p-value = 0.5584
```

```
fit.thyroid <- aov(thyroid ~ group, data = data.thyroid)
summary(fit.thyroid)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        2   2744  1372.1   3.623 0.0445 *
## Residuals   21   7953   378.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("No difference in variance. Significant difference in average
    level for each group.")
```

```
## No difference in variance. Significant difference in average
##     level for each group.
```

2. (Page 258, 2)

```
data.rat <- data.frame(type = rep(c('A', 'B', 'C'), each = 5) %>% factor(),
                       group = rep(1:5, 3) %>% factor(),
                       hours = c(1.16, 2.11, 1.82, 1.41, 0.51, 1.30, 3.28, 4.98,
                                 2.59, 0.59, 3.36, 5.28, 4.81, 2.04, 5.05))

fit.rat <- aov(hours ~ type + group, data= data.rat)
summary(fit.rat)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## type         2  18.45   9.224   7.204 0.0162 *
## group        4  10.72   2.680   2.093 0.1736
## Residuals    8  10.24   1.280
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("No significant difference for each group but there is significant
    difference for various types of drug. ")
```

```
## No significant difference for each group but there is significant
##     difference for various types of drug.
```

3. (Page 258, 3)

```
dummy <- data.frame(before = c(4, 3.5, 3.2, 3.2, 3.3, 3.4, 2.7, 4.8, 4.5, 3.8),
                    after = c(5.4, 4.7, 5.2, 4.8, 4.6, 4.9, 3.8, 6.1, 5.9, 4.9))
dummy_diff <- with(dummy, after - before)

pills <- data.frame(before = c(3.5, 3.3, 3.2, 4.5, 4.3, 3.2, 4.2, 5., 4.3, 3.6),
                    after = c(4.7, 4.4, 4., 5.2, 5., 4.3, 5.1, 6.5, 4., 4.7))
pills_diff <- with(pills, after - before)

dt.combined <- data.frame(dummy = dummy_diff, pills = pills_diff) %>%
    pivot_longer(cols = everything(), names_to = "group", values_to = "hours") %>%
    mutate(group = factor(group))
bartlett.test(hours ~ group, data = dt.combined)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  hours by group
## Bartlett's K-squared = 2.7758, df = 1, p-value = 0.0957
```

```
t.test(dummy_diff, pills_diff, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  dummy_diff and pills_diff
## t = 2.9203, df = 18, p-value = 0.009137
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1430897 0.8769103
## sample estimates:
## mean of x mean of y
##      1.39      0.88
```

```
cat("No difference in variance. By t-test we can conclude that there is significant
    different between the effect of pills and that of placebo.")
```

```
## No difference in variance. By t-test we can conclude that there is significant
##     different between the effect of pills and that of placebo.
```

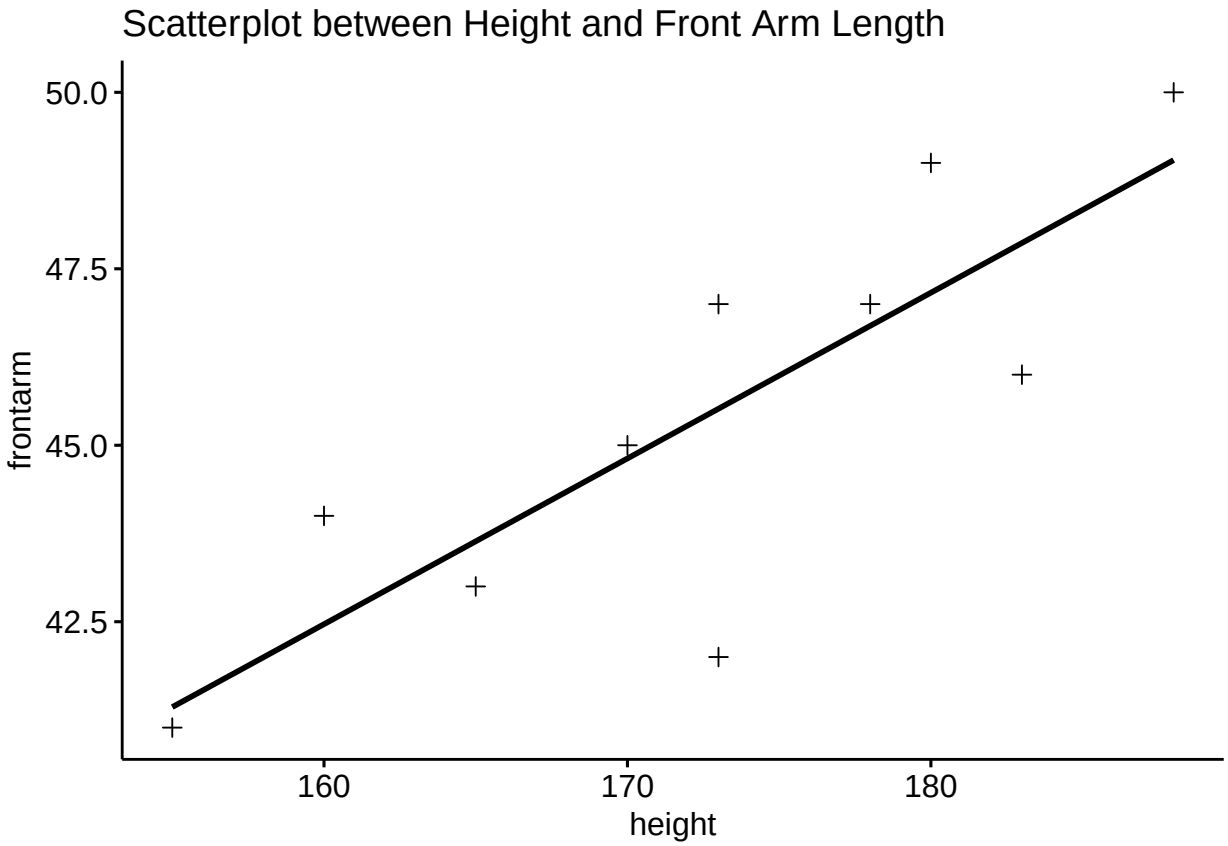
2.5 Solutions for chapter 6

1. (Page 260 1)

(1)

```
data.cstu <- data.frame(height= c(170, 173, 160, 155, 173, 188, 178, 183, 180, 165),
                        frontarm = c(45, 42, 44, 41, 47, 50, 47, 46, 49, 43))
ggscatter(data = data.cstu, x = 'height', y = 'frontarm', palette = 'npg',
          shape = 3, title = 'Scatterplot between Height and Front Arm Length',
          add = 'reg.line')
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
with(data.cstu, cor(height, frontarm))
```

```
## [1] 0.8227162
```

```
with(data.cstu, cor.test(height, frontarm))
```

```
##
##  Pearson's product-moment correlation
##
## data:  height and frontarm
## t = 4.0936, df = 8, p-value = 0.003468
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4006045 0.9567450
## sample estimates:
##       cor
## 0.8227162
```

Figure shows they are jointly normal distributed. We use Pearson's test for correlation ($P < 0.01$). The results show that Height and Length of front arm are correlated, with Pearson's correlation coefficient as 0.82.

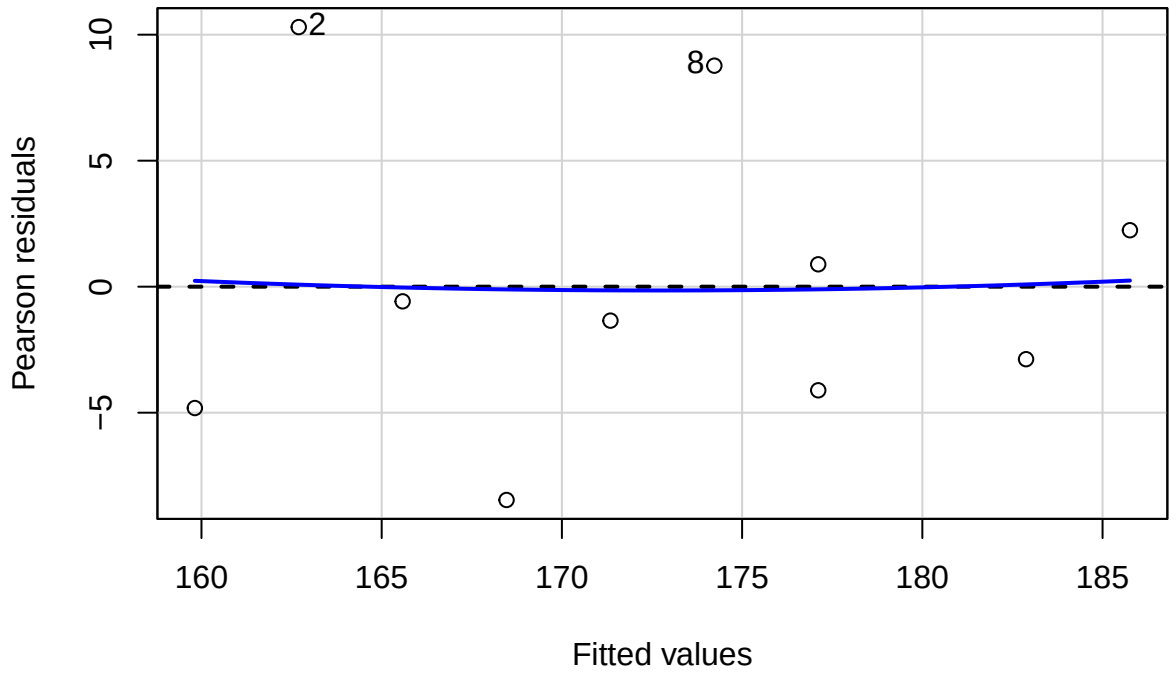
(2)

```
mod.xy <- lm(height ~ frontarm, data = data.cstu)
summary(mod.xy)
```

```
##
## Call:
## lm(formula = height ~ frontarm, data = data.cstu)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4643 -3.8036 -0.9643  1.9018 10.3010
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.6276    32.0311   1.300  0.22993
## frontarm      2.8827     0.7042   4.094  0.00347 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.235 on 8 degrees of freedom
## Multiple R-squared:  0.6769, Adjusted R-squared:  0.6365
## F-statistic: 16.76 on 1 and 8 DF,  p-value: 0.003468
```

```
residualPlot(mod.xy, id = TRUE)
```




```
# aov(mod.xy)
```

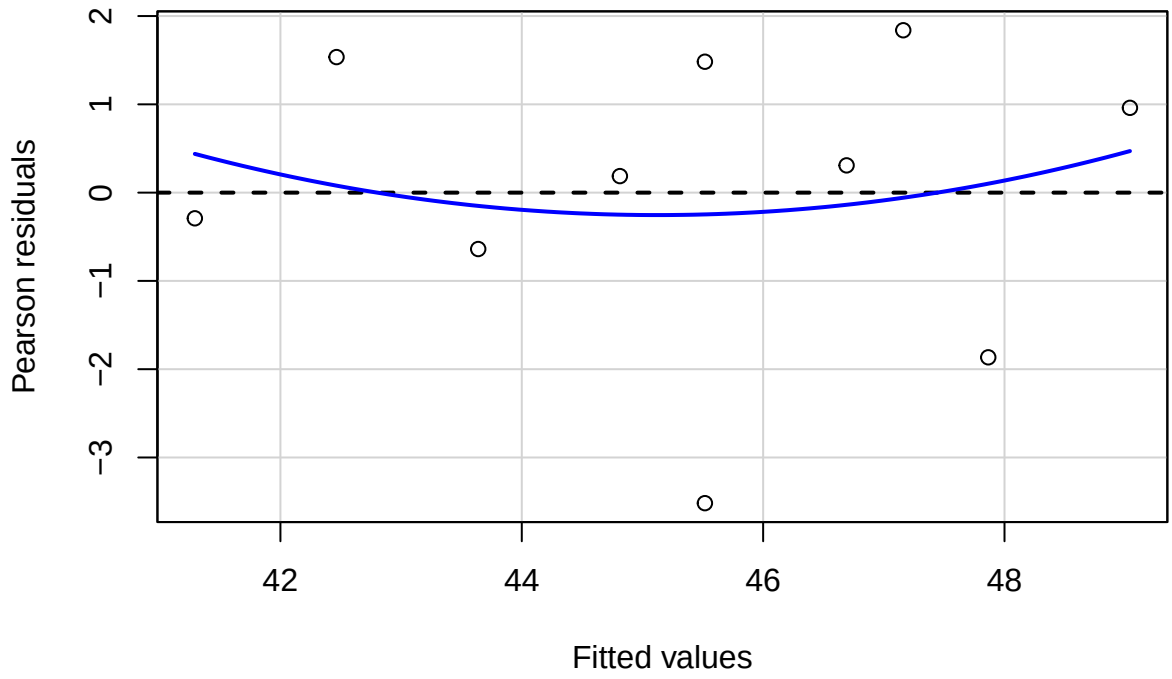
The residual plot shows the residuals follow the linear assumption, and the estimated coefficient is significantly different with 0.

```
mod.yx <- lm(frontarm ~ height, data = data.cstu)
summary(mod.yx)
```

```
##
## Call:
## lm(formula = frontarm ~ height, data = data.cstu)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5174 -0.5519  0.2478  1.3521  1.8390
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.89610    9.91053   0.494  0.63456
## height       0.23481    0.05736   4.094  0.00347 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.78 on 8 degrees of freedom
## Multiple R-squared:  0.6769, Adjusted R-squared:  0.6365
## F-statistic: 16.76 on 1 and 8 DF,  p-value: 0.003468
```

```
residualPlot(mod.yx)
```




```
# aov(mod.yx)
```

The residual plot shows the residuals don't follow linear assumption, we should make transformation to variables.

2. (Page 260 3(1))

Omitted.

3 Remarks

Here lists several notes that might be forgotten

3.1 chapter 2: Numerical variables, categorical variables and corresponding descriptive tables and figures

- CV(coefficient variation) is the ratio of sample standard deviation with sample mean ($\frac{S}{\bar{X}} \times 100\%$)
- To test the normality of data, use `qqnorm(x); qqline(x)` for figuring and `shapiro.test(x)` for numerical testing
- When comparing two groups of different total numbers, please standardize them implicitly(SMR) or explicitly
- Ubiquitous statistical figures include **barplot, pie, lines, hist, stem and boxplot, .etc**
- Statistical figures with standard errors can be plotted with `arrows(x0 = x, y0 = y - sd, x1 = x, y1 = y +sd, angle = , length = )`

3.2 chapter 3: Estimation and hypothesis testing for overall mean

- Standard error: the standard deviation of sample mean, denoted by $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- Estimation: moment estimation and maximizing the likelihood(mle)
- General methods for deriving interval estimation: (1) find out a random variable $f(X_1, \ldots, X_n, \theta)$ with a given distribution having nothing with θ, denoted as $F(x)$; (2) find out a, b so that $F(b) - F(a) = 1 - \alpha$; (3) transform $f(X_1, \ldots, X_n, \theta) \in (a, b]$ into the form of θ that $\theta \in (\theta_1, \theta_2)$
- Summarize the basic approaches of interval estimation: (1) if σ is given, then $z-$distribution; (2) if σ is not given, then $t-$distribution
- Hypothesis testing. For variance testing, (1) $H_0 : \sigma = \sigma_0$, use χ^2-test; (2) $H_0 : \sigma_1^2 = \sigma_2^2$, use $F-$test. For sample mean, similar with interval estimation in single sample testing. For two samples, (1) paired, use `t.test(x, y, paired = TRUE)`; (2) assuming equal variance, use `t.test(x, y, var.equal = TRUE)`; else use `t.test(x, y, var.equal = FALSE)`
- Understand two type errors. (Under fixed type 1 error, we can reduce type 2 error by putting on samples)
- Non-parametric testing for sample median: Rank-sum test, use `wilcox.test(x, y)`.

3.3 chapter 4: Binomial, Poisson distribution and hypothesis testing for categorical variables

- Hypothesis testing for poisson distributed samples: (1) single sample, use normal approximation $\frac{X - \lambda_0}{\sqrt{\lambda_0}} \sim N(0, 1)$; (2) two samples, under $H_0 : \lambda_1 = \lambda_2$ if $X_1 \sim \text{Pois}(n_1\lambda_1)$ and $X_2 \sim \text{Pois}(n_2\lambda_2)$, then $(X_1/n_1 - X_2/n_2)/\sqrt{X_1/n_1^2 + X_2/n_2^2} \approx N(0, 1)$
- In following situations χ^2-test works, (1) Goodness-of-fit, `chisq.test(x, p = )`; (2) test for non-correlation, `chisq.test(xmatrix, correct = )`; or u can use fisher exact testing, `fisher.test(data, alternative =)`; (3) paired design, `mcnemar.test(data)`

3.4 chapter 5: Analysis of variance

- Major goal: compare the mean between two or more groups
- Basic assumptions: (1) within each group, samples follow normal distribution with identical variance (2) independence
- R code for anova: (1) testing the homoscedasticity, `barlett.test(score ~ group, data = )`; (2) anova, `score.aov = aov(score ~ group, data = ); summary(score.aov)`; (3) randomized-block design, `aov(score ~ blocknum + group, data = ) %>% summary()`; (4) analysis of covariance, `fit <- aov(score ~ age + group, data = ); summary(fit)`; to show group effects, `library(effects); effect("group", fit)`; to pairwise compare, `library(multcomp); res.vsl = glht(fit, linfct = mcp(group = c("a -b = 0", "b - c = 0", "a - c = 0"))); summary(res.vsl, test = adjusted("bonferroni"))`

- For a pairwise comparison, we can use a post hoc test with adjust p-value, `pairwise.t.test(group, score, p.adjust.method = "holm")`

3.5 chapter 6: regression analysis

- Compute the correlations. (1) when data follows joint normal distribution, then use pearson correlation `cor(x, y, method = 'pearson'); cor.test(..)`; (draw figures to verify the joint normality first) (2) when data does not follow normal distribution, then use spearman's correlation, `cor(x, y, method = "spearman")`
- Statistical inference in linear regression model. (1) t-test for coefficients, `confint(model, level = .95)`; (2) for fitting and prediction, the interval estimations are not identical, `predict(model, newdata = data.frame(var = ), interval = c('prediction', 'confidence'), level = .95)`.
- Life table (Omitted)

4 Suggestions for advanced R or statistics

For more advanced usages of R code, you can refer to

- A tutorial written by Dongfeng Li, Peking University, attached link: `https://www.math.pku.edu.cn/teachers/lidf/docs/Rbook/html/_Rbook/index.html`
- R Cookbook, 2nd Edition
- R Graphics Cookbook, 2nd Edition (for making figures)

To obtain more knowledge on statistics, u are welcomed to take the course hosted by Dr. Jia on next semester, namely *Statistical modeling* !