2024 01 08 VB-STA5 Exam in statistics - Solution Guide

1. Math and Statistics exam performance

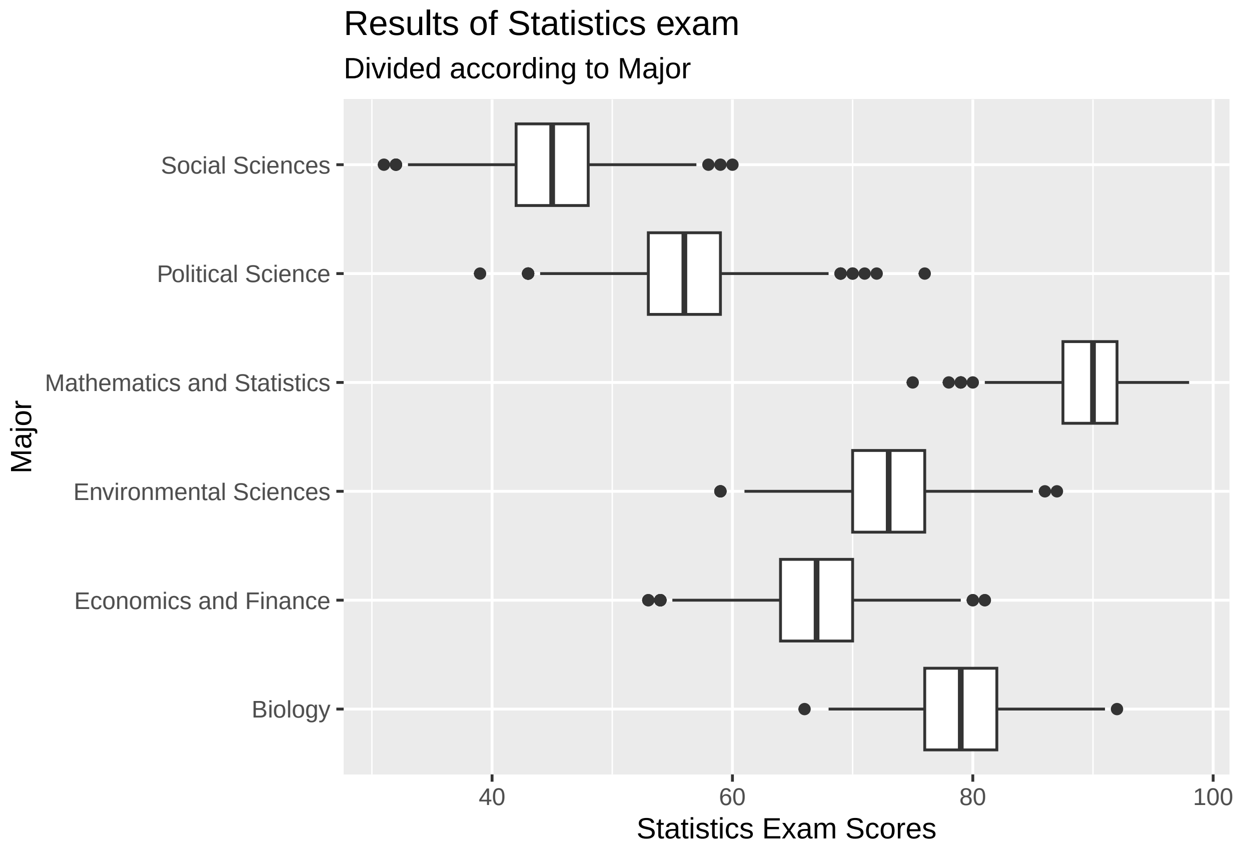
Dataset *data/students_exam_performance.csv* contains information about students that participated both in Mathematics and Statistics class.

a) Recreate the plot:

```
students <- readr::read_csv('data/students_exam_performance.csv')
```

```
## Rows: 4891 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): semester, major, minor, math_grade, statistics_grade
## dbl (3): stud.id, math_score, statistics_score
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ggplot(students) +
  geom_boxplot(aes(y = major, x = statistics_score))+
  labs(x = 'Statistics Exam Scores',
       y = 'Major',
       title = 'Results of Statistics exam',
       subtitle = 'Divided according to Major')
```



b) Describe the plot.

- boxplot presenting statistics exam score according to the students major
- mean score per major are in order (from lowest to highest) Social Sciences, Political Science, Economics and Finance, Environmental Sciences, Biology and last Mathematics and Statistics.
- social sciences lowest score ranging from 30-60 points.
- ... for the rest of majors
- Mathematics and Statistics highest score ranging from 75 to 97-8. Also with narrowest distribution. Except for few outliers students scored in 80-100 points range.

c) Check whether there is a significant difference between a Mathematics Exam score for *Economics and Finance* major students with minor in *Mathematics and Statistics*, and *Economics and Finance* major students with other minors. Conduct an appropriate test for this situation.

Difference of means t-test.

$H_0 : \mu_{m\_minor_{m}s} - \mu_{m\_minor_{n}ot_{m}s} = 0$

$H_A : \mu_{m\_minor_{m}s} - \mu_{m\_minor_{n}ot_{m}s} = 0 \neq 0$

H0: There is no difference between mean Math exam score for *Economics and Finance* major students with minor in *Mathematics and Statistics*, and *Economics and Finance* major students with other minors

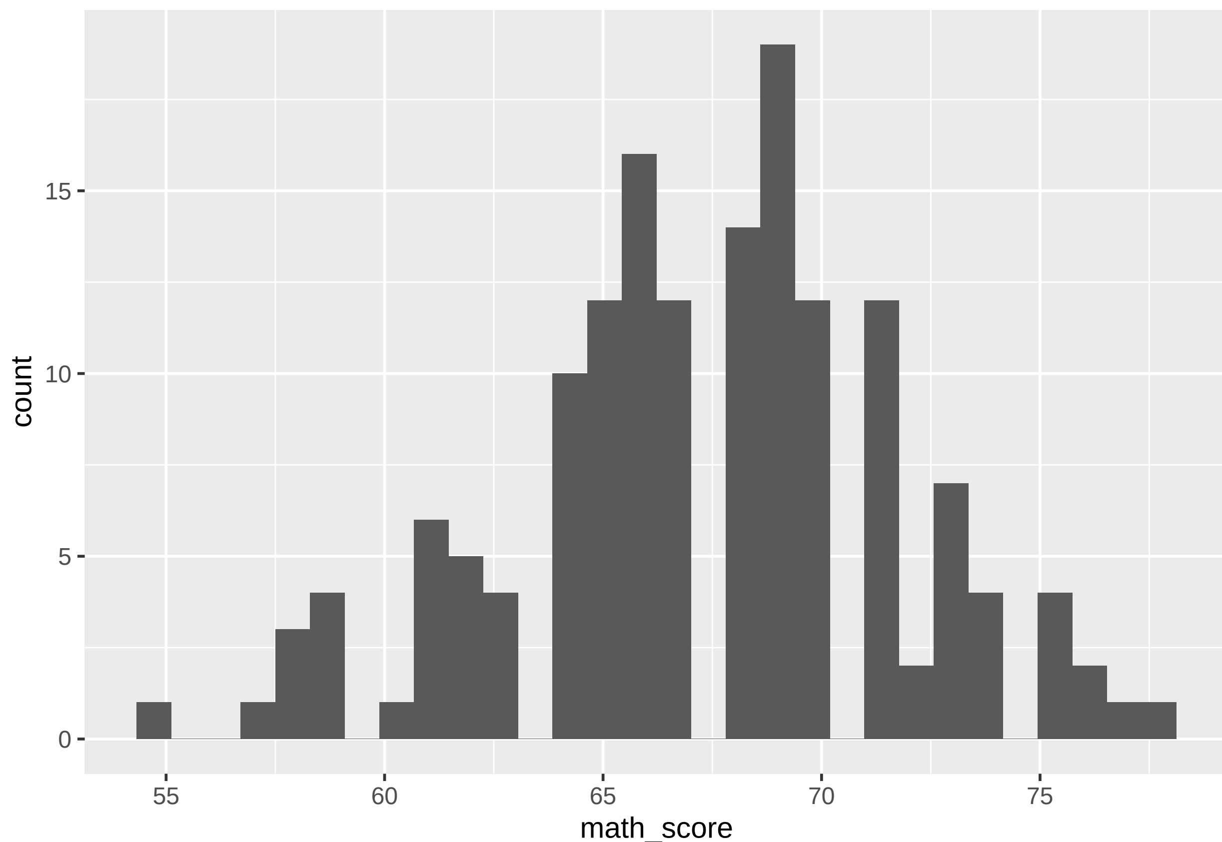
HA: There is difference between mean Math exam score for *Economics and Finance* major students with minor in *Mathematics and Statistics*, and *Economics and Finance* major students with other minors alpha significance level - 0.05

onditions check:

Normality:

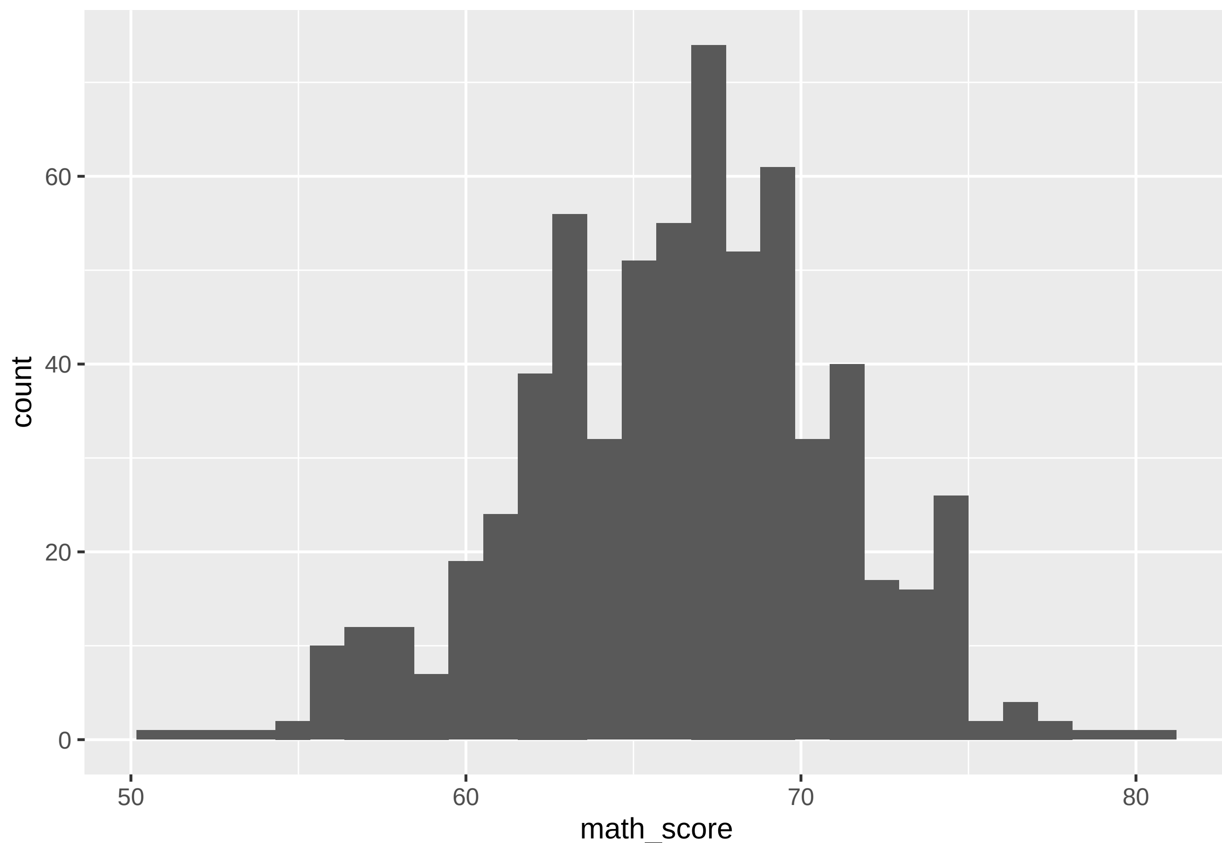
```r
students_b_m <- students %>%
  filter(major == 'Economics and Finance') %>%
  filter(minor == 'Mathematics and Statistics')
students_b_nm <- students %>%
  filter(major == 'Economics and Finance') %>%
  filter(minor != 'Mathematics and Statistics')

ggplot(students_b_m) +
  geom_histogram(aes(x = math_score))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
ggplot(students_b_nm) +
  geom_histogram(aes(x = math_score))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The variables distributions look normal.

We assume that observations are independent.

- short version

```r
t.test(students_b_m$math_score, students_b_nm$math_score)
```

```
##
##  Welch Two Sample t-test
##
## data:  students_b_m$math_score and students_b_nm$math_score
## t = 2.8193, df = 238.75, p-value = 0.005217
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3347965 1.8878868
## sample estimates:
## mean of x mean of y
##  67.33987  66.22853
```

p-value is smaller than alpha significance level, thus we reject null hypothesis in favour of the alternative. There is statistically significant difference between mean Math exam score for *Economics and Finance* major students with minor in *Mathematics and Statistics*, and *Economics and Finance* major students with other minors

- long version

```r
(point_estimate <- mean(students_b_m$math_score) -
   mean(students_b_nm$math_score))
```

```
## [1] 1.111342
```

```r
(nrow(students_b_m))
```

```
## [1] 153
```

```r
(nrow(students_b_nm))
```

```
## [1] 652
```

```r
dof <- 152
```

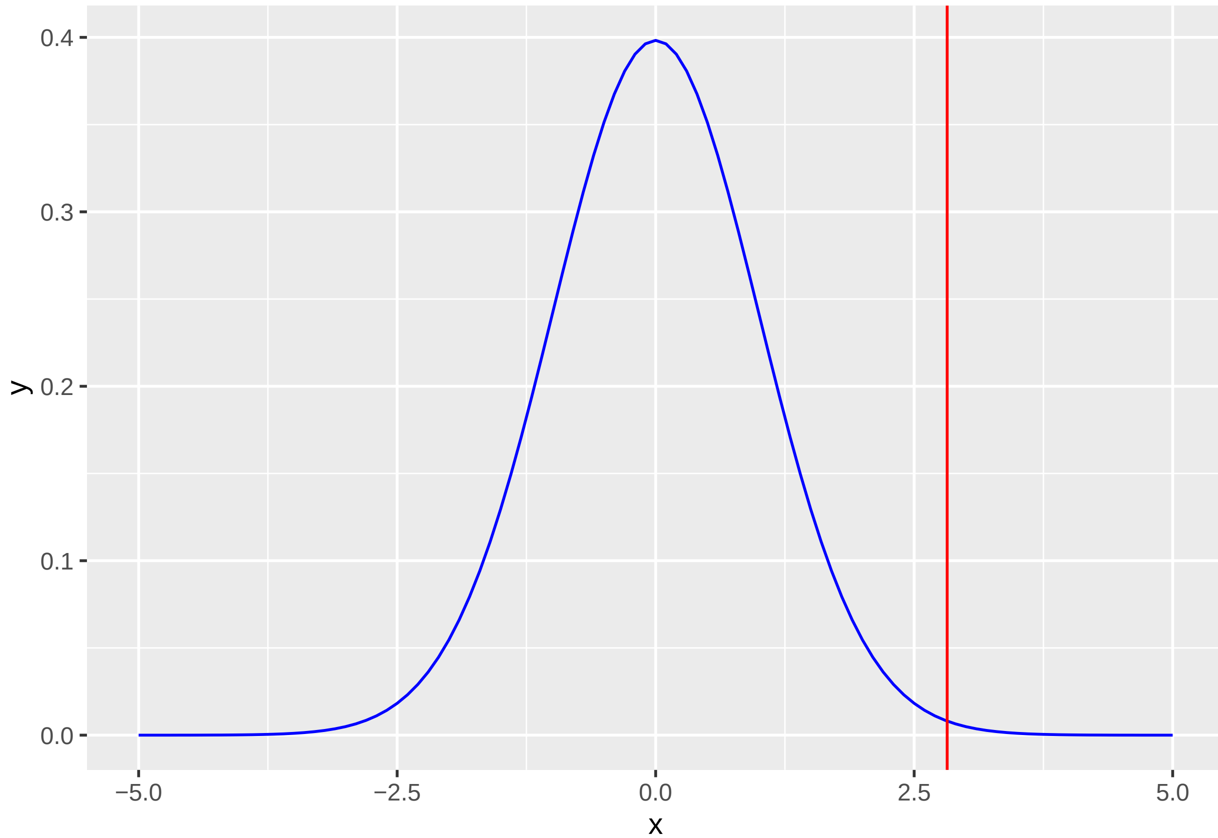
```r
(SE <- sqrt((sd(students_b_m$math_score)^2/nrow(students_b_m)) +
              (sd(students_b_nm$math_score)^2/nrow(students_b_nm))))
```

```
## [1] 0.3941954
```

```r
(t_score <- (point_estimate - 0)/SE)
```

```
## [1] 2.819266
```

```r
ggplot(data.frame(x = seq(-5, 5, length=100)), aes(x = x)) +
  stat_function(fun = dt, args = list(df = dof), color = 'blue') +
  geom_vline(aes(xintercept = t_score),  color = 'red')
```




```
(p_value <- 2 * (1- pt(t_score, df = dof)))
```

```
## [1] 0.005454935
```

p-value is smaller than alpha significance level, thus we reject null hypothesis in favour of the alternative. There is statistically significant difference between mean Math exam score for *Economics and Finance* major students with minor in *Mathematics and Statistics*, and *Economics and Finance* major students with other minors

Childs seatbelt - car seat legislation

```
accidents <- readr::read_csv("data/CarSeatLegislation.csv")
```

```
## Rows: 33258 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Restraint, Injury
## dbl (1): ID
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

a) Present the number and proportion of 'No Injury' accidents divided according to the implemented protection.

```
accidents %>%
  filter(Injury == 'No Injury') %>%
  group_by(Restraint) %>%
  tally() %>%
  mutate(Proportion = n/sum(n)) %>%
  knitr::kable()
```

Restraint	n	Proportion
Car Seat	1532	0.2977070
Lap and Shoulder Belt	974	0.1892732
Lap-Only Belt	871	0.1692577
No Restraint	1769	0.3437621

b) Is there a correlation in between type of Injury and Implemented protection? Form hypothesis, check for conditions, and conduct a statistical test.

```
(two_way_table <- accidents %>%
  group_by(Injury, Restraint) %>%
  tally() %>%
  ungroup() %>%
  spread(Restraint, n))
```

```
## # A tibble: 5 x 5
##   Injury        `Car Seat` Lap and Shoulder Bel~1 `Lap-Only Belt` `No Restraint`
##   <chr>              <int>                  <int>           <int>          <int>
## 1 Fatal               1241                    978             772           6201
## 2 Incapacitati~       1136                   1088            1103           6645
## 3 No Injury           1532                    974             871           1769
## 4 Non-incapaci~       1610                   1233            1190            468
## 5 Possible Inj~       1111                    772             683           1881
## # i abbreviated name: 1: `Lap and Shoulder Belt`
```

Chi square test for independence.

Conditions for the test:

- dataset is independent
- expected cases should be more than 5

H0: There is no correlation in between type of injury and Implemented protection in car accidents where children were passengers.

H0: There is correlation in between type of injury and Implemented protection in car accidents where children were passengers.

alpha significance level - 0.05

```r
colnames(two_way_table)
```

```
## [1] "Injury"                 "Car Seat"               "Lap and Shoulder Belt"
## [4] "Lap-Only Belt"          "No Restraint"
```

```r
sum_all <- sum(two_way_table$`Car Seat`) +
  sum(two_way_table$`Lap and Shoulder Belt`) +
  sum(two_way_table$`Lap-Only Belt`) +
  sum(two_way_table$`No Restraint`)

two_way_table %>% mutate(CS_exp = (`Car Seat` +
                                     `Lap and Shoulder Belt` +
                                     `Lap-Only Belt` +
                                     `No Restraint`)*sum(two_way_table$`Car Seat`)/sum_all) %>%
  mutate(LSB_exp = (`Car Seat` +
                      `Lap and Shoulder Belt` +
                      `Lap-Only Belt` +
                      `No Restraint`)*sum(two_way_table$`Lap and Shoulder Belt`)/sum_all) %>%
  mutate(LOB_exp = (`Car Seat` +
                      `Lap and Shoulder Belt` +
                      `Lap-Only Belt` +
                      `No Restraint`)*sum(two_way_table$`Lap-Only Belt`)/sum_all) %>%
  mutate(NR_exp = (`Car Seat` +
                     `Lap and Shoulder Belt` +
                     `Lap-Only Belt` +
                     `No Restraint`)*sum(two_way_table$`No Restraint`)/sum_all)
```

```
## # A tibble: 5 x 9
##   Injury `Car Seat` Lap and Shoulder Bel~1 `Lap-Only Belt` `No Restraint` CS_exp
##   <chr>       <int>                  <int>           <int>          <int>  <dbl>
## 1 Fatal        1241                    978             772           6201  1832.
## 2 Incap~       1136                   1088            1103           6645  1988.
## 3 No In~       1532                    974             871           1769  1026.
## 4 Non-i~       1610                   1233            1190            468   897.
## 5 Possi~       1111                    772             683           1881   887.
## # i abbreviated name: 1: `Lap and Shoulder Belt`
## # i 3 more variables: LSB_exp <dbl>, LOB_exp <dbl>, NR_exp <dbl>
```

All expected values are above 5.

- short version

```r
two_way_table %>% select(-1) %>%
chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 5751.9, df = 12, p-value < 2.2e-16
```

We reject null hypothesis in favour of the alternative. There is correlation in between type of injury and Implemented protection in car accidents where children were passengers.

- long version

```
two_way_table <- two_way_table %>%
  mutate(CS_exp = (`Car Seat` +
                   `Lap and Shoulder Belt` +
                   `Lap-Only Belt` +
                   `No Restraint`)*sum(two_way_table$`Car Seat`)/sum_all) %>%
  mutate(LSB_exp = (`Car Seat` +
                    `Lap and Shoulder Belt` +
                    `Lap-Only Belt` +
                    `No Restraint`)*sum(two_way_table$`Lap and Shoulder Belt`)/sum_all) %>%
  mutate(LOB_exp = (`Car Seat` +
                    `Lap and Shoulder Belt` +
                    `Lap-Only Belt` +
                    `No Restraint`)*sum(two_way_table$`Lap-Only Belt`)/sum_all) %>%
  mutate(NR_exp = (`Car Seat` +
                   `Lap and Shoulder Belt` +
                   `Lap-Only Belt` +
                   `No Restraint`)*sum(two_way_table$`No Restraint`)/sum_all)
```

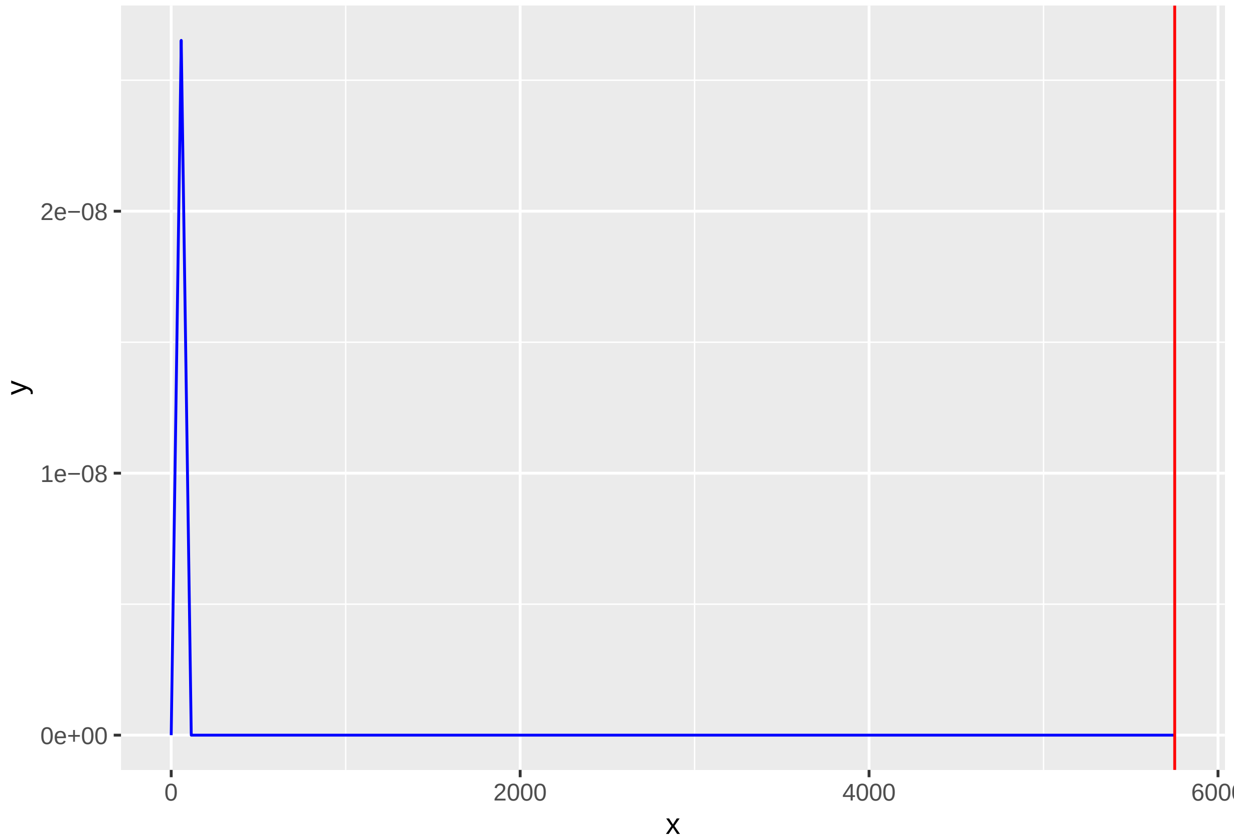
```
(chi2_stat <- sum(((two_way_table$`Car Seat` - two_way_table$CS_exp)/ two_way_table$CS_exp^0.5)^2) +
   sum(((two_way_table$`Lap and Shoulder Belt` - two_way_table$LSB_exp)/ two_way_table$LSB_exp^0.5)^2) +
   sum(((two_way_table$`Lap-Only Belt` - two_way_table$LOB_exp)/ two_way_table$LOB_exp^0.5)^2) +
   sum(((two_way_table$`No Restraint` - two_way_table$NR_exp)/ two_way_table$NR_exp^0.5)^2))
```

```
## [1] 5751.894
```

```
(dof <- 4*3)
```

```
## [1] 12
```

```
ggplot(data.frame(x = seq(0, 25, length=100)), aes(x = x)) +
  stat_function(fun = dchisq, args = list(df = dof), color = 'blue') +
  geom_vline(aes(xintercept = chi2_stat),  color = 'red')
```




```
(p_value <- 1 - pchisq(chi2_stat, df = dof))
```

```
## [1] 0
```

We reject null hypothesis in favour of the alternative. There is correlation in between type of injury and Implemented protection in car accidents where children were passengers.

3.Wild blueberries yield prediction

Three datasets about wild blueberry farming are provided:

- *data/blueberries_insects.csv* contains information about pollinating insects presence
- *data/blueberries_weather.csv* contains information about weather (temperature and rain)
- *data/blueberries_yield.csv* contains information about size of the fruit, seeds, and final yield.

a) Join all three datasets.

```
blue_insects <- readr::read_csv("data/blueberries_insects.csv")
```

```
## Rows: 777 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (5): plotID, honeybee, bumbles, andrena, osmia
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
blue_weather <- readr::read_csv("data/blueberries_weather.csv")
```

```
## Rows: 777 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (6): plotID, MaxTemp, MinTemp, AverageTemp, RainingDays, AverageRainingDays
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
blue_size <- readr::read_csv("data/blueberries_yield.csv")
```

```
## Rows: 777 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (6): plotID, clonesize, fruitset, fruitmass, seeds, yield
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
blue <- blue_insects %>% left_join(blue_size, by = join_by(plotID)) %>% left_join(blue_weather, by = joi
```

b) Which variables have statistically significant influence on the blueberry yield? Create multiple regression model and tune it.

```r
fit <- lm(yield ~ clonesize+
            honeybee+
            bumbles+
            andrena+
            osmia+
            MaxTemp+
            MinTemp+
            AverageTemp+
            # RainingDays+ 1st and only removed
            AverageRainingDays+
            fruitset+
            fruitmass+
            seeds,
          data = blue)
```

```r
summary(fit)
```

```
##
## Call:
## lm(formula = yield ~ clonesize + honeybee + bumbles + andrena +
##     osmia + MaxTemp + MinTemp + AverageTemp + AverageRainingDays +
##     fruitset + fruitmass + seeds, data = blue)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -532.49  -75.86    1.34   69.04  457.73
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -12670.617   3242.694  -3.907 0.000102 ***
## clonesize              -8.068      1.217  -6.632 6.27e-11 ***
## honeybee               50.382      8.460   5.955 3.96e-09 ***
## bumbles               232.006     99.022   2.343 0.019387 *
## andrena               354.130     31.204  11.349  < 2e-16 ***
## osmia                 557.956     35.445  15.741  < 2e-16 ***
## MaxTemp             -3357.676    878.935  -3.820 0.000144 ***
## MinTemp             -4052.480   1062.080  -3.816 0.000147 ***
## AverageTemp          7213.292   1894.474   3.808 0.000152 ***
## AverageRainingDays   -937.701     47.900 -19.576  < 2e-16 ***
## fruitset             8835.374    497.078  17.775  < 2e-16 ***
## fruitmass          -26593.008   3037.508  -8.755  < 2e-16 ***
## seeds                 349.116     24.115  14.477  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 123.9 on 764 degrees of freedom
## Multiple R-squared:  0.9918, Adjusted R-squared:  0.9917
## F-statistic:  7688 on 12 and 764 DF,  p-value: < 2.2e-16
```

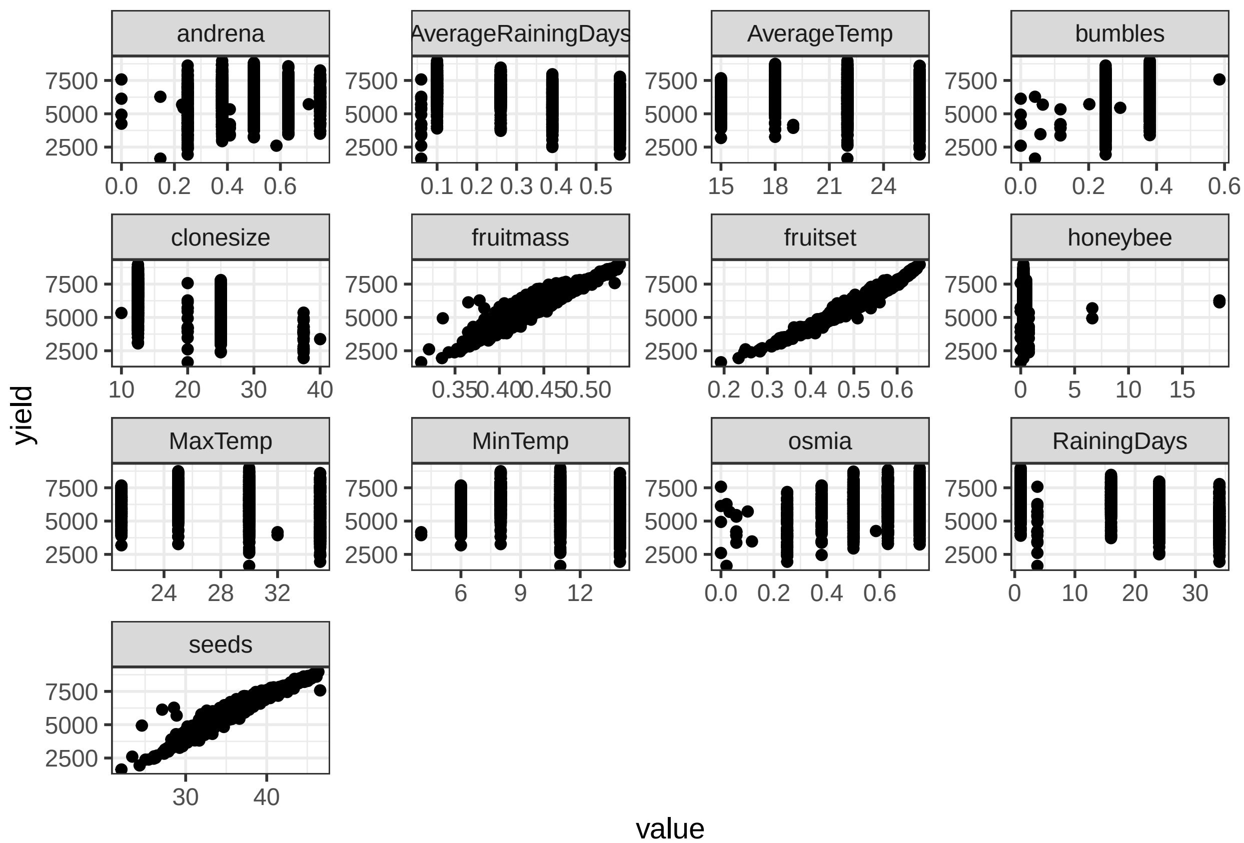
yield = -12670.617 + clonesize* (-8.068) + honeybee * 50.382 + bumbles * 232.006 + andrena * 354.130 + osmia * 557.956 + MaxTemp * -3357.676 + MinTemp * -4052.480 + AverageTemp * 7213.292 + AverageRainingDays * -937.701 + fruitset * 8835.374 + fruitmass * -26593.008 + seeds * 349.116

Following variables have statistically significant influence on the bluberry yield: clonesize, honeybee, bumbles, andrena, osmia, MaxTemp, MinTemp, AverageTemp, AverageRainingDays, fruitset, fruitmass, seeds.

c) What should be satisfied for model (3b) to be valid. Check if the model you created is valid?

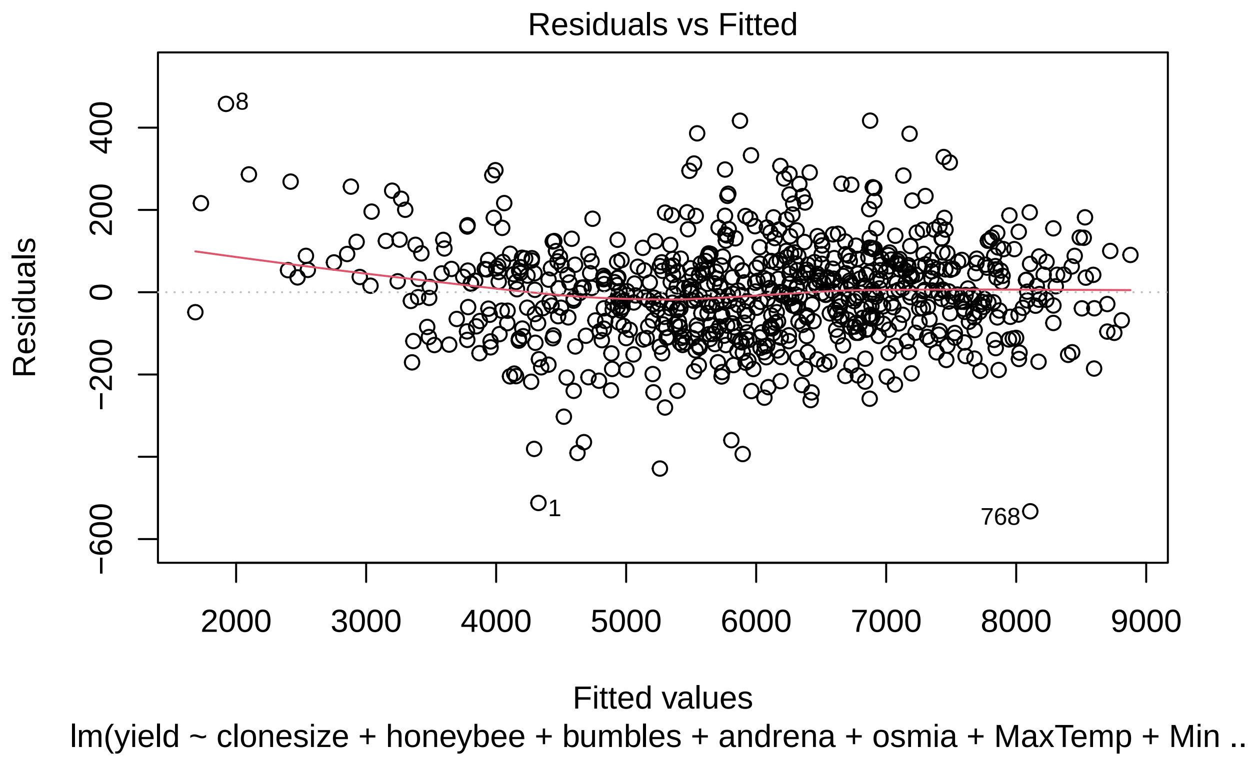
- linearity
- nearly normal residuals
- constant variability
- independent observations

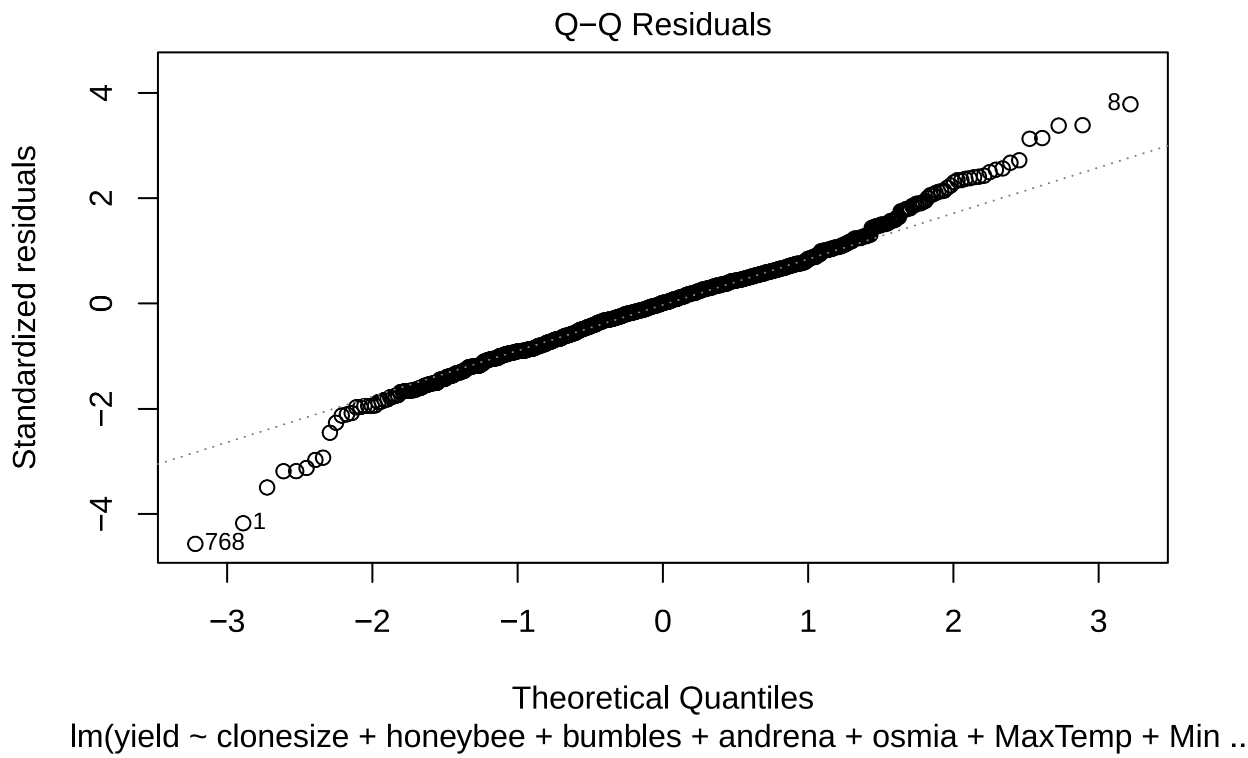
```
blue %>% select(-1) %>%
  gather(-yield, key = "var", value = "value") %>%
  ggplot(aes(x = value, y = yield, )) +
    geom_point() +
    facet_wrap(~ var, scales = "free") +
    theme_bw()
```

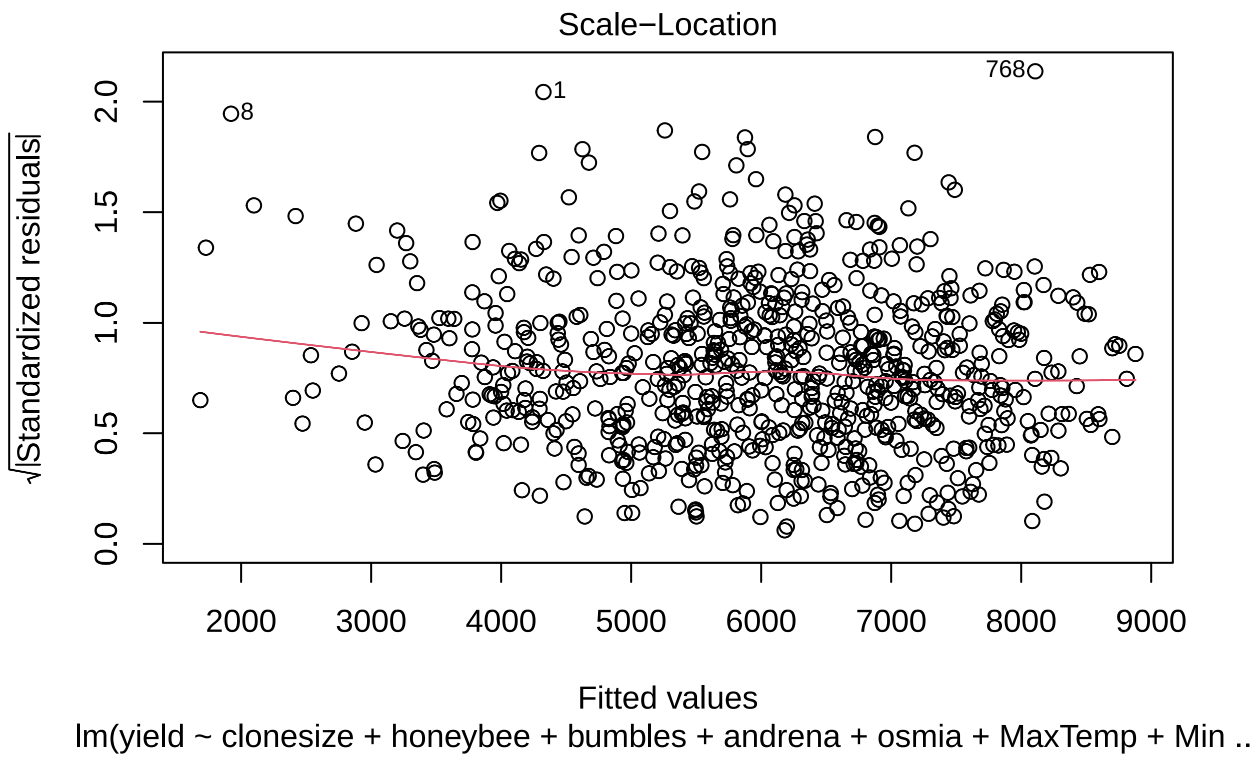


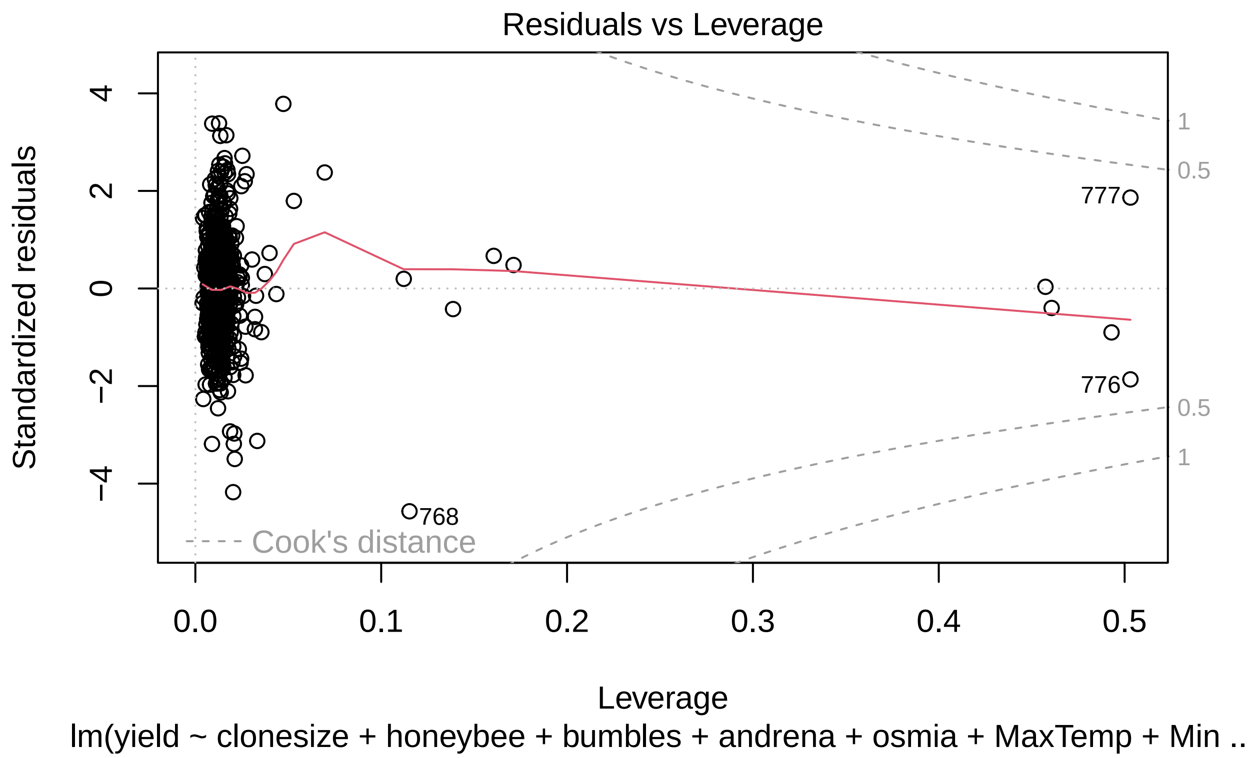
All numerical variables, with continous scale have clear linear trend. ‘AverageTemp’ has a more quadratic function tendency, however it’s hard to pinpoint.

```
plot(fit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(yield ~ clonesize + honeybee + bumbles + andrena + osmia + MaxTemp + Min ..
400
200
0
−200
−600
2000
3000
4000
5000
6000
7000
8000
9000
8
1
768


Q−Q Residuals
Standardized residuals
4
2
0
−2
−4
Theoretical Quantiles
−3
−2
−1
0
1
2
3
8
1
768
lm(yield ~ clonesize + honeybee + bumbles + andrena + osmia + MaxTemp + Min ..


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2000
3000
4000
5000
6000
7000
8000
9000
Fitted values
lm(yield ~ clonesize + honeybee + bumbles + andrena + osmia + MaxTemp + Min ..
8
1
768



No trend visible in the first plot. There seems to be consant variability to residuals. q-q plot also suggests that there is normal distribution to residuals.

We assume observations are independent.

d) Construct confidence interval for multiplication parameter of 'seeds' variable.

We are 95% confident that 'seeds' multiplier value for linear regression fit is between:

```
summary(fit)$coefficients[13] - 1.96 * summary(fit)$coefficients[26]
```

```
## [1] 301.8514
```

and

```
summary(fit)$coefficients[13] + 1.96 * summary(fit)$coefficients[26]
```

```
## [1] 396.3814
```