2024 02 21 VB-STA5 Reexam in Statistics - Solution Guide

Wednesday 21st of Fabruary.

The exam set consists of 3 main exercises with 9 sub exercises in total.

Each sub exercise is weighted equally when grading the hand ins.

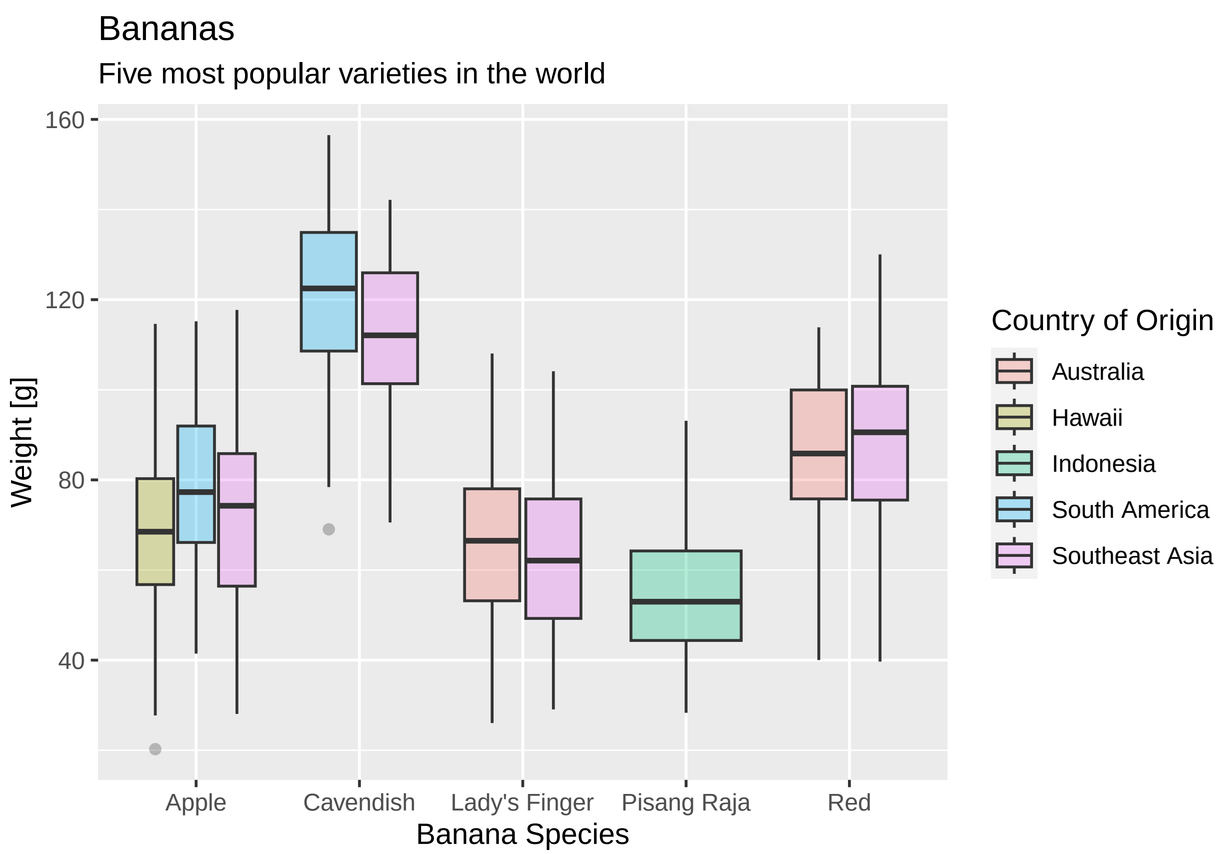
1. Bananas

Dataset *data/bananas_dataset.csv* contains information about five most popular banana varieties in the world.

a) Recreate the plot:

```
bananas <- readr::read_csv('data/banana_dataset.csv')

bananas %>%
  ggplot() +
  geom_boxplot(aes(Species, Weight_g, fill = Origin), alpha = 0.3)+
  labs(x = 'Banana Species',
       y = 'Weight [g]',
       title = 'Bananas',
       fill = 'Country of Origin',
       subtitle = 'Five most popular varieties in the world')
```



b) Describe the plot. Include description of what is presented, how it is presented, is there grouping, and what information about those groups can be read.

c) For bananas grown in Southeast Asia, present the average length and average weight divided according to Species.

```
bananas %>% filter(Origin == 'Southeast Asia') %>%
  group_by(Species) %>%
  summarize(`Average Length [cm]` = mean(Length_cm),
            `Average Weight [g]` = mean(Weight_g)) %>%
  knitr::kable()
```

Species	Average Length [cm]	Average Weight [g]
Apple	13.10311	72.87696
Cavendish	19.12166	112.81263
Lady's Finger	11.85156	62.76676
Red	15.04770	87.87070

d) Banana plant originated in Southeast Asia, however Ecuador is the biggest producer of bananas in the world. Select a relevant statistical test and use it to test whether average Ecuadorian (South American) *Cavendish* bananas are bigger/heavier than average Southeast Asia *Cavendish* bananas. Form hypothesis, check for conditions, conduct a statistical test, and form conclusions.

Difference of means t-test.

$H_0 : \mu_{m\_south_america} - \mu_{m\_southeast_asia} = 0$

$H_A : \mu_{m\_south_america} - \mu_{m\_southeast_asia} > 0$

H0: There is no difference between mean *South American Cavendish banana* weight, and mean *Southeast Asia Cavendish banana* weight

HA: There is difference between mean *South American Cavendish banana* weight, and mean *Southeast Asia Cavendish banana* weight, and mean *South American Cavendish banana* is heavier.

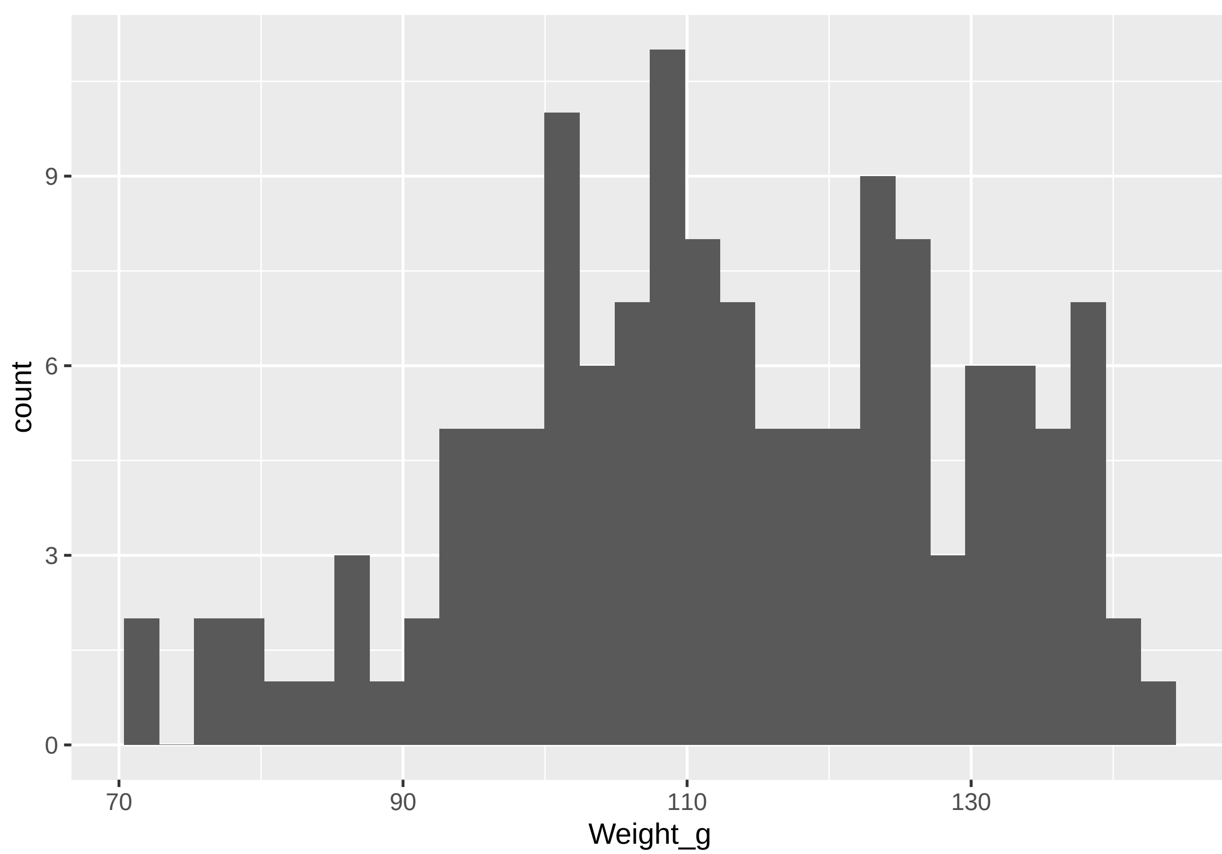
alpha significance level - 0.05

conditions check:

Normality:

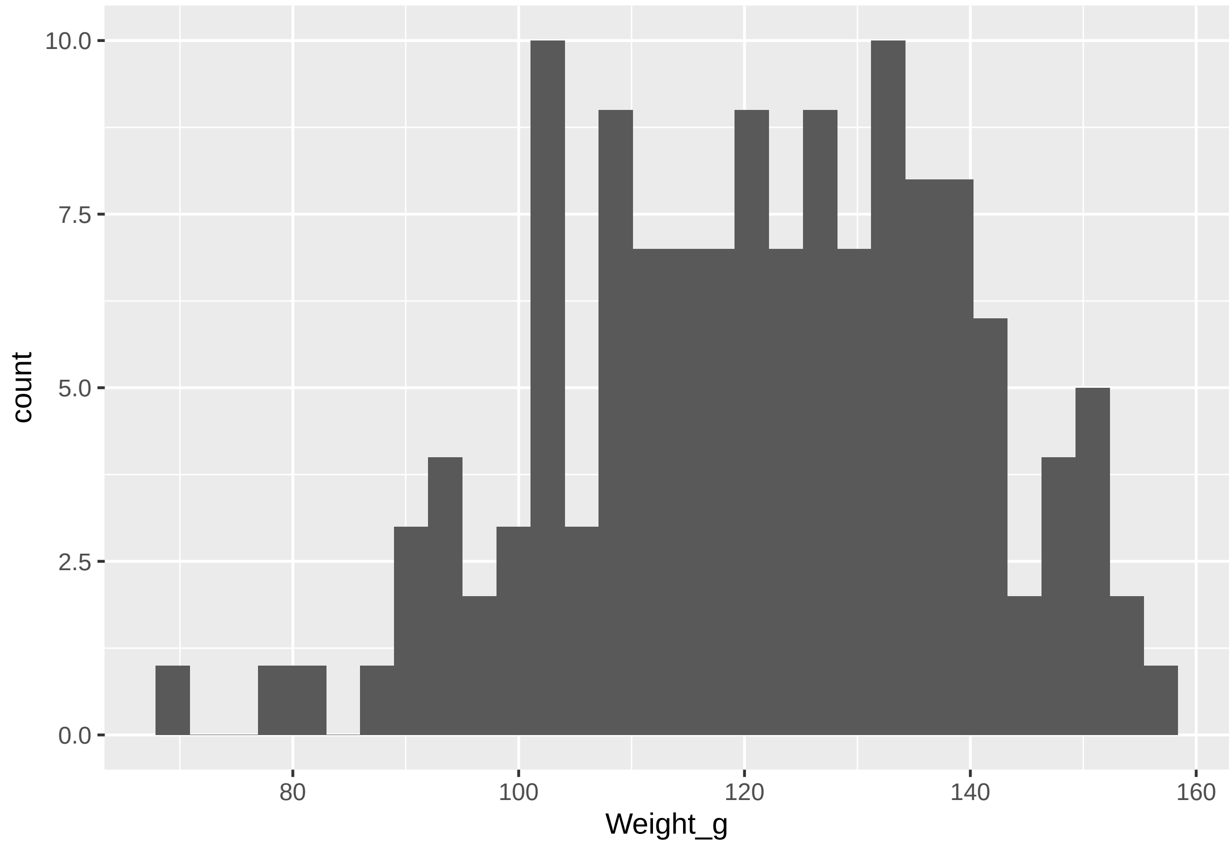
```
bananas %>% filter(Origin == 'Southeast Asia') %>%
  filter(Species == 'Cavendish') %>%
  ggplot() +
  geom_histogram(aes(x = Weight_g))
```

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




```
bananas %>% filter(Origin == 'South America') %>%
  filter(Species == 'Cavendish') %>%
  ggplot() +
  geom_histogram(aes(x = Weight_g))
```

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



The variables distributions look normal.

We assume that observations are independent.

- short version

```r
bananas %>% filter(Origin %in% c('South America', 'Southeast Asia')) %>%
  filter(Species == 'Cavendish') %>%
t.test(Weight_g~Origin, data = ., alternative = 'greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  Weight_g by Origin
## t = 4.1655, df = 272.69, p-value = 2.087e-05
## alternative hypothesis: true difference in means between group South America and group Southeast Asia
## 95 percent confidence interval:
##  5.169005      Inf
## sample estimates:
##  mean in group South America mean in group Southeast Asia
##                     121.3737                     112.8126
```

p-value is smaller than alpha significance level, thus we reject null hypothesis in favour of the alternative. There is difference between mean *South American Cavendish banana* weight, and mean *Southeast Asia Cavendish banana* weight, and mean *South American Cavendish banana* is heavier.

- long version

```
SE_Asia <- bananas %>% filter(Origin == 'Southeast Asia') %>%
  filter(Species == 'Cavendish')
S_America <- bananas %>% filter(Origin == 'South America') %>%
  filter(Species == 'Cavendish')

(point_estimate <- mean(S_America$Weight_g) -
   mean(SE_Asia$Weight_g))
```

```
## [1] 8.561112
```

```
(nrow(S_America))
```

```
## [1] 137
```

```
(nrow(SE_Asia))
```

```
## [1] 140
```

```
(dof <- 136)
```

```
## [1] 136
```

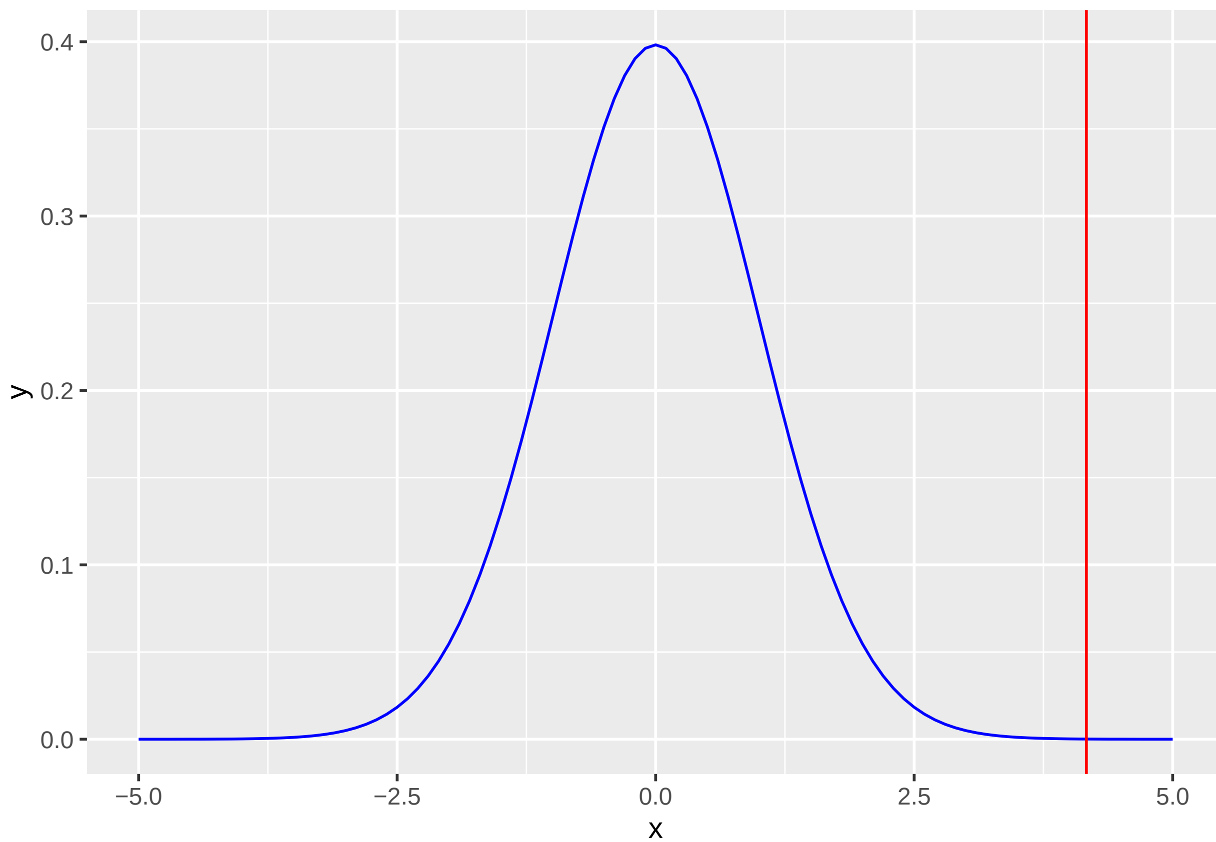
```
(SE <- sqrt((sd(S_America$Weight_g)^2/nrow(S_America)) +
              (sd(SE_Asia$Weight_g)^2/nrow(SE_Asia))))
```

```
## [1] 2.055249
```

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

```
## [1] 4.165488
```

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

```
## [1] 2.744795e-05
```

p-value is smaller than alpha significance level, thus we reject null hypothesis in favour of the alternative. here is difference between mean *South American Cavendish banana* weight, and mean *Southeast Asia Cavendish banana* weight, and mean *South American Cavendish banana* is heavier.

2. USA population in 2020

The dataset *data/US_state_capitol_population_2020.csv* contains information about population and race in state capitols in US. Population of entire USA in 2020 is summarized in a table below:

```
population <- readr::read_csv('data/US_state_capitol_population_2020.csv')
```

```
## Rows: 7 Columns: 52
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Race
## dbl (51): Montgomery, Juneau, Phoenix, Little Rock, Sacramento, Denver, Hart...
##
## 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.
```

```r
whole_population <- sum(population$USA)
population %>% select(Race, USA) %>% mutate(proportion = USA/whole_population)
```

```
## # A tibble: 7 x 3
##   Race                                            USA proportion
##   <chr>                                         <dbl>      <dbl>
## 1 White                                     195223627    0.575
## 2 Black or African American                  45077102    0.133
## 3 American Indian and Alaska Native           4308841    0.0127
## 4 Asian                                      20881305    0.0615
## 5 Native Hawaiian and Other Pacific Islander   994348    0.00293
## 6 Two or More Races                           9943478    0.0293
## 7 Hispanic or Latino                         63306813    0.186
```

a) Select a relevant statistical test, and use it to check, whether Race distribution of Atlanta population is following the same distribution as Race distribution of the entire country. Form hypothesis, check for conditions, conduct a statistical test, and form conclusions.

Chi square test for goodness of fit.

Conditions for the test:

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

H0: The population of Atlanta follows racial distribution of the entire USA.

H0: The racial distribution of Atlanta's population is statistically significantly different from racial distribution of the entire USA.

alpha significance level - 0.05

Check for expected values:

```r
Atlanta_all <- sum(population$Atlanta)
(pop_Atlanta <- population %>% select(Race, USA, Atlanta) %>%
  mutate(proportion = USA/whole_population) %>%
  mutate(expected = proportion*Atlanta_all))
```

```
## # A tibble: 7 x 5
##   Race                                         USA Atlanta proportion expected
##   <chr>                                      <dbl>   <dbl>      <dbl>    <dbl>
## 1 White                                     1.95e8  193003    0.575    308646.
## 2 Black or African American                 4.51e7  256340    0.133     71266.
## 3 American Indian and Alaska Native         4.31e6     997    0.0127     6812.
## 4 Asian                                     2.09e7   22941    0.0615    33013.
## 5 Native Hawaiian and Other Pacific Islander 9.94e5     499    0.00293    1572.
## 6 Two or More Races                         9.94e6   13465    0.0293    15720.
## 7 Hispanic or Latino                        6.33e7   49872    0.186    100087.
```

All expected values are above 5.

Short version

- short version

```r
chisq.test(pop_Atlanta$Atlanta, p=pop_Atlanta$proportion, rescale.p = FALSE)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  pop_Atlanta$Atlanta
## X-squared = 558240, df = 6, p-value < 2.2e-16
```

We reject null hypothesis in favour of the alternative. The racial distribution of Atlanta's population is statistically significantly different from racial distribution of the entire USA.

- long version

```r
(pop_Atlanta <- pop_Atlanta %>%
  mutate(Z = (Atlanta - expected)/sqrt(expected)) %>%
  mutate(Z2 = Z^2))
```

```
## # A tibble: 7 x 7
##   Race                              USA Atlanta proportion expected      Z     Z2
##   <chr>                           <dbl>   <dbl>      <dbl>    <dbl>  <dbl>  <dbl>
## 1 White                          1.95e8  193003    0.575    308646. -208.  4.33e4
## 2 Black or African American      4.51e7  256340    0.133     71266.  693.  4.81e5
## 3 American Indian and Alaska N~ 4.31e6     997    0.0127     6812.  -70.5 4.96e3
## 4 Asian                          2.09e7   22941    0.0615    33013.  -55.4 3.07e3
## 5 Native Hawaiian and Other Pa~ 9.94e5     499    0.00293    1572.  -27.1 7.32e2
## 6 Two or More Races              9.94e6   13465    0.0293    15720.  -18.0 3.24e2
## 7 Hispanic or Latino             6.33e7   49872    0.186    100087. -159.  2.52e4
```

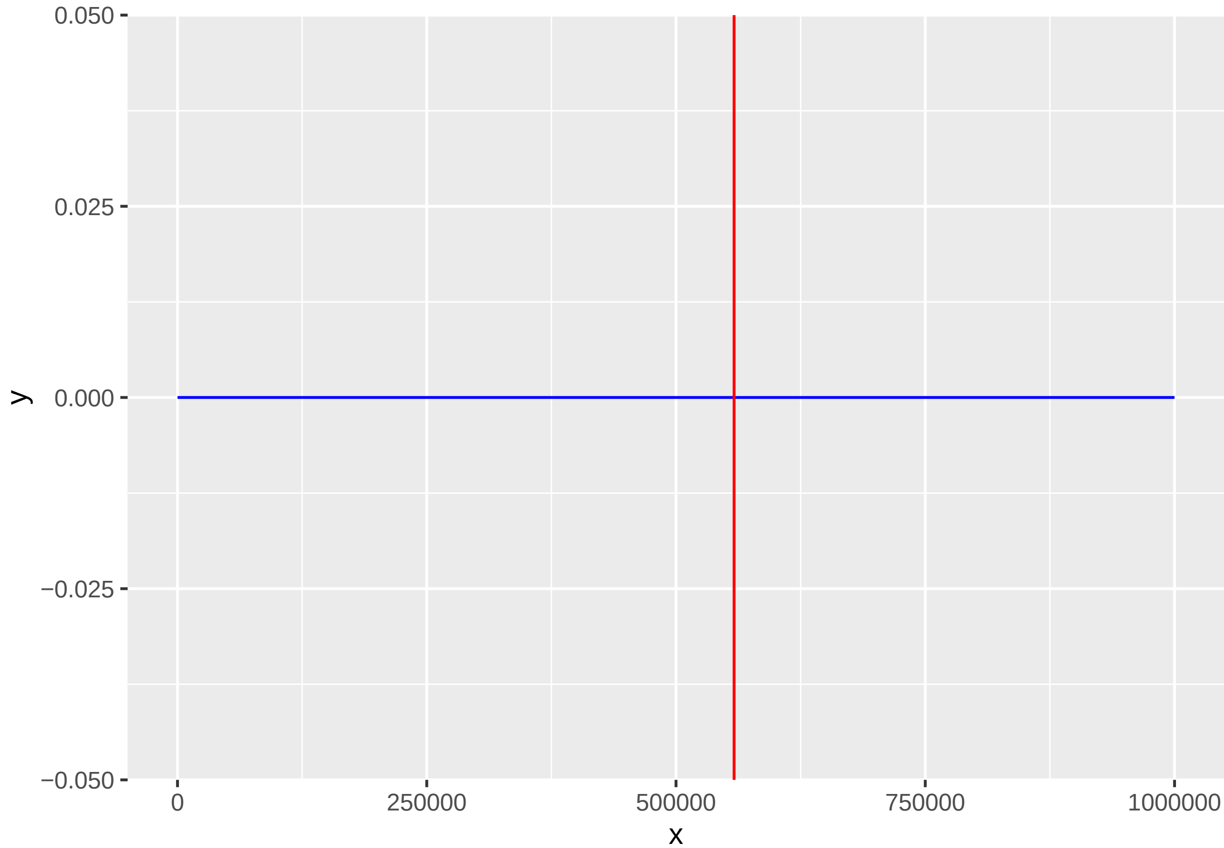
```r
(chi2_stat <- sum(pop_Atlanta$Z2))
```

```
## [1] 558239.8
```

```r
(dof <- 7-1)
```

```
## [1] 6
```

```r
ggplot(data.frame(x = seq(0, 1000000, 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. The racial distribution of Atlanta's population is statistically significantly different from racial distribution of the entire USA.

3. Airfares

airq412.csv contains information about airfares and passengers for the U.S. Domestic Routes for 4th quarter of 2012. Norwegian Airlines wants to break into the U.S. market with a new route in between Point Place, Wisconsin (-) and Los Angeles, California (LAX). Currently there are **no commercial** flights from Point Place, and due to that the city is not included in the database.

The distance in between two cities is 2260 miles and is expected to have approximately 150 passengers per week.

a) Create a linear model to predict 'Average Fare' using 'Distance'.

```
airfares <- readr::read_csv('data/airq412.csv')
```

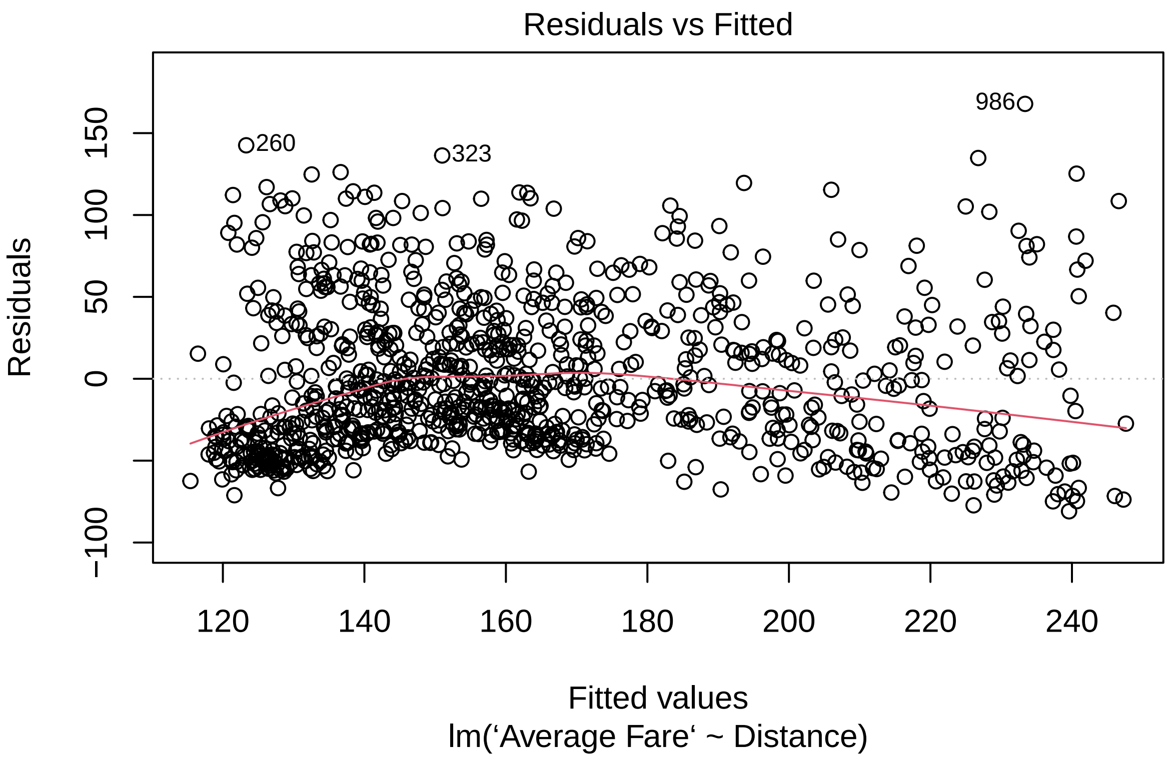
```r
fit <- lm(`Average Fare` ~ Distance, data = airfares)
summary(fit)
```

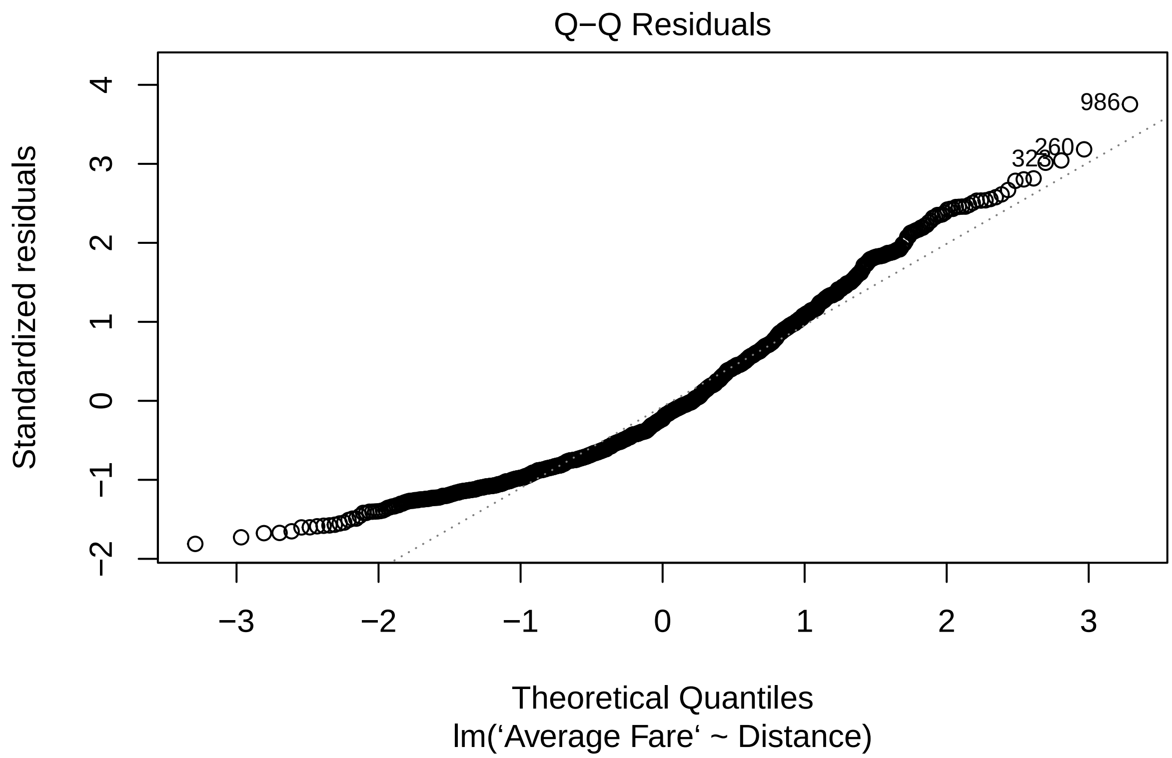
```
##
## Call:
## lm(formula = `Average Fare` ~ Distance, data = airfares)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -80.92 -34.45 -10.28  27.85 167.85
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.100e+02  2.729e+00   40.30   <2e-16 ***
## Distance    5.054e-02  2.206e-03   22.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.84 on 998 degrees of freedom
## Multiple R-squared:  0.3448, Adjusted R-squared:  0.3441
## F-statistic: 525.1 on 1 and 998 DF,  p-value: < 2.2e-16
```

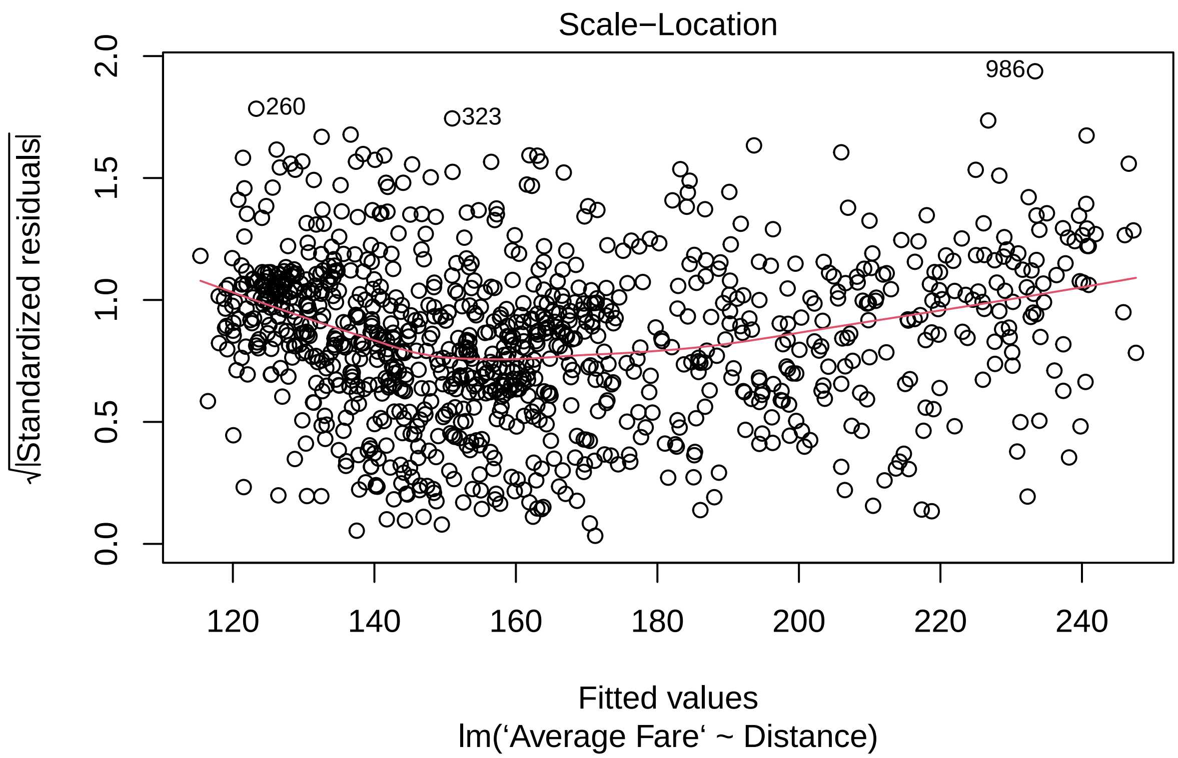
$$AverageFare = 0.05054214 \cdot Distance + 109.9537$$

b) Evaluate the model from exercise 3a (check if conditions are fulfilled).

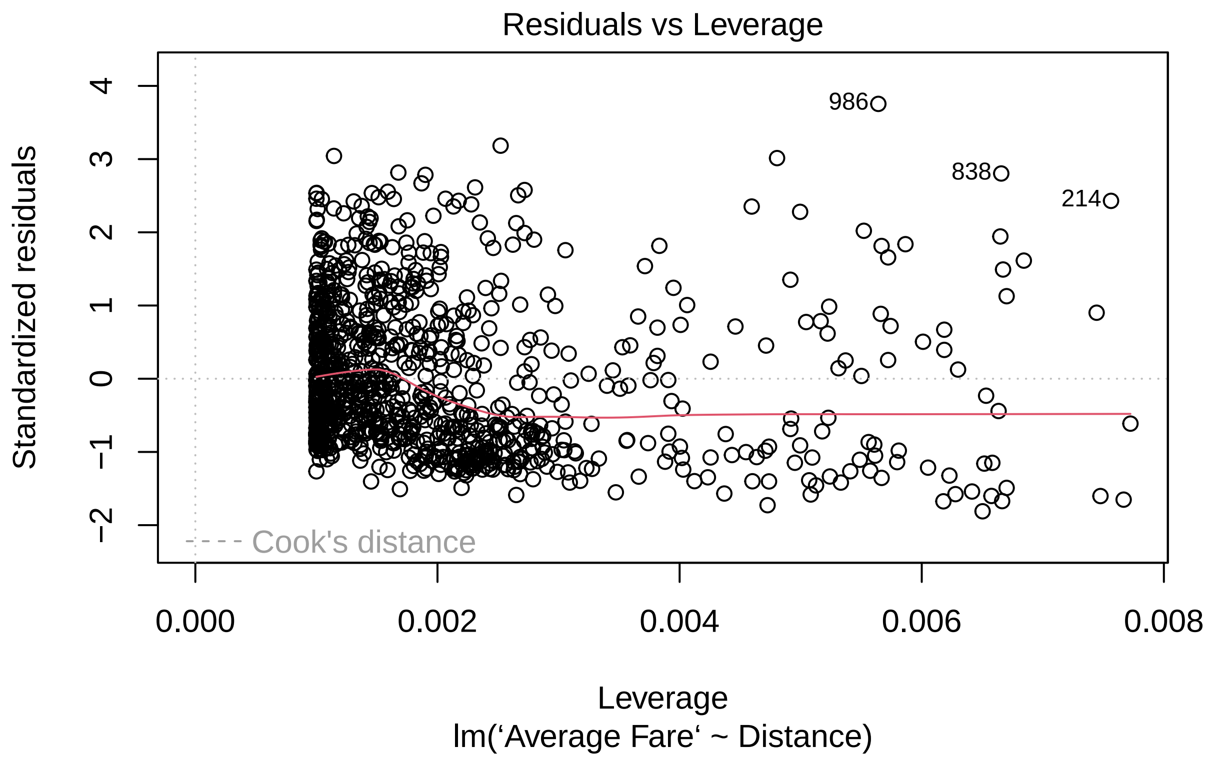
```r
plot(fit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(‘Average Fare‘ ~ Distance)
986
260
323
150
100
50
0
−100
120
140
160
180
200
220
240


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(‘Average Fare‘ ~ Distance)
986
260
323
−3
−2
−1
0
1
2
3
4

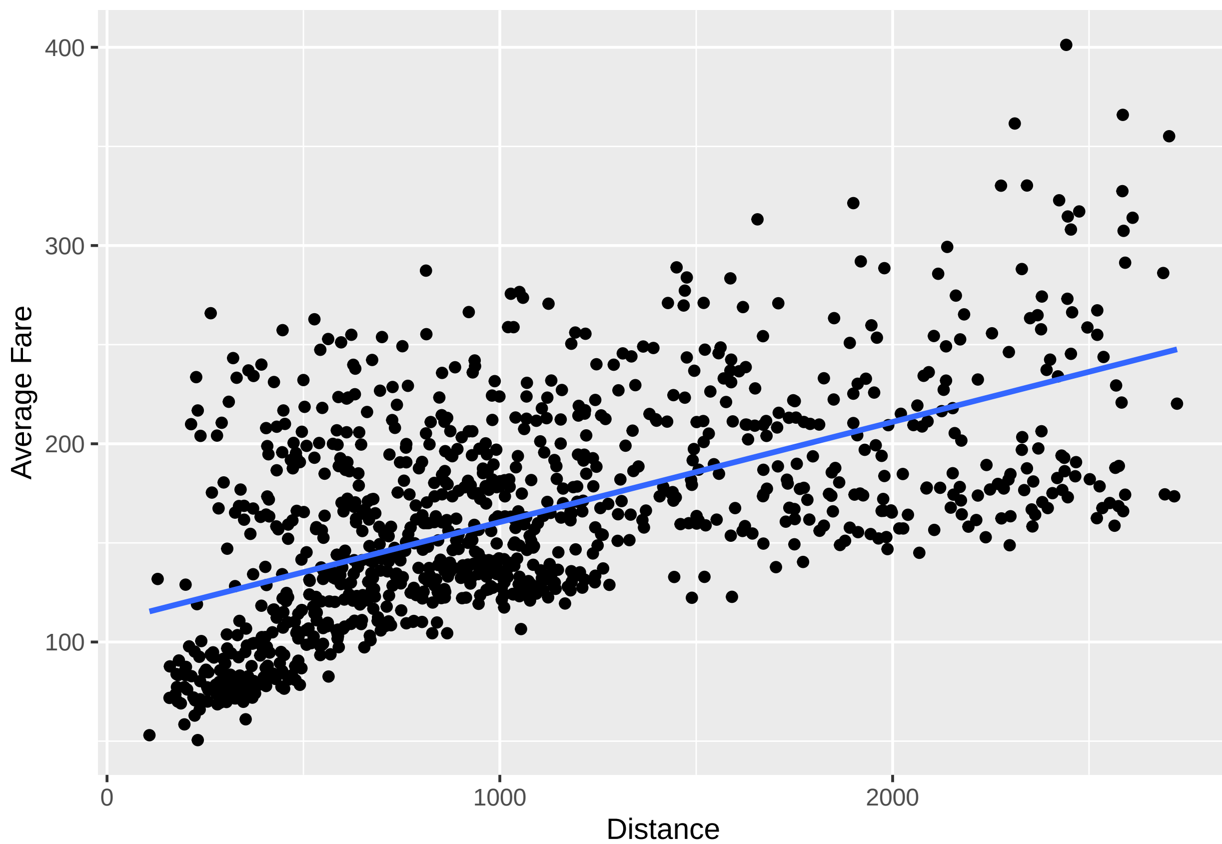


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
120
140
160
180
200
220
240
Fitted values
lm(‘Average Fare‘ ~ Distance)
260
323
986




```
ggplot(airfares) +
  geom_point(aes(Distance, `Average Fare`)) +
  geom_smooth(aes(Distance, `Average Fare`), method = lm, se = FALSE)
```

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



- Linear trend

There seems to be linear trend in data.

- Constant variability of residuals and normal distributed residuals

In the first plot we can see that residuals are not so evenly distributed and in the second we can see nearly normal distribution of residuals with some divergence in tails. It seems that a linear model might not be the best solution here.

- Independent observations

The cases might not be independent as each airport might have different starting price.

c) Propose a price for a ticket from Point Place to Los Angeles using the model created in exercise 3a.

Proposed price:

```
fit$coefficients[1] + 2260*fit$coefficients[2]
```

```
## (Intercept)
##     224.179
```

d) Create two multiple regression models to predict *Average Fare* using:

- *Distance*, *Average Weekly Passengers*, *Market Share MLA*,
- *Distance*, *Market Share MLA*

Which one is better in your opinion and why? Use chosen model to predict an average fare for new route between Point Place and Los Angeles.

1st model

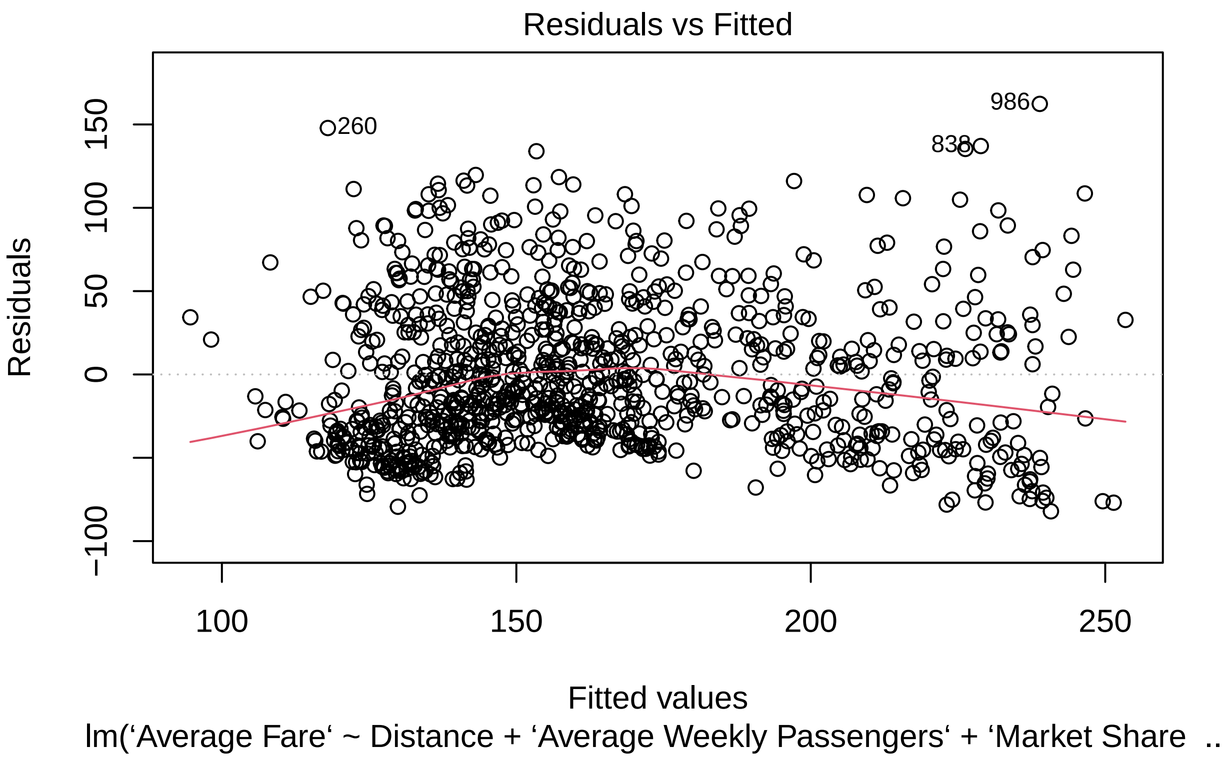
```r
fit <- lm(`Average Fare` ~ Distance +
            `Average Weekly Passengers` +
            `Market Share MLA`, data = airfares)
```

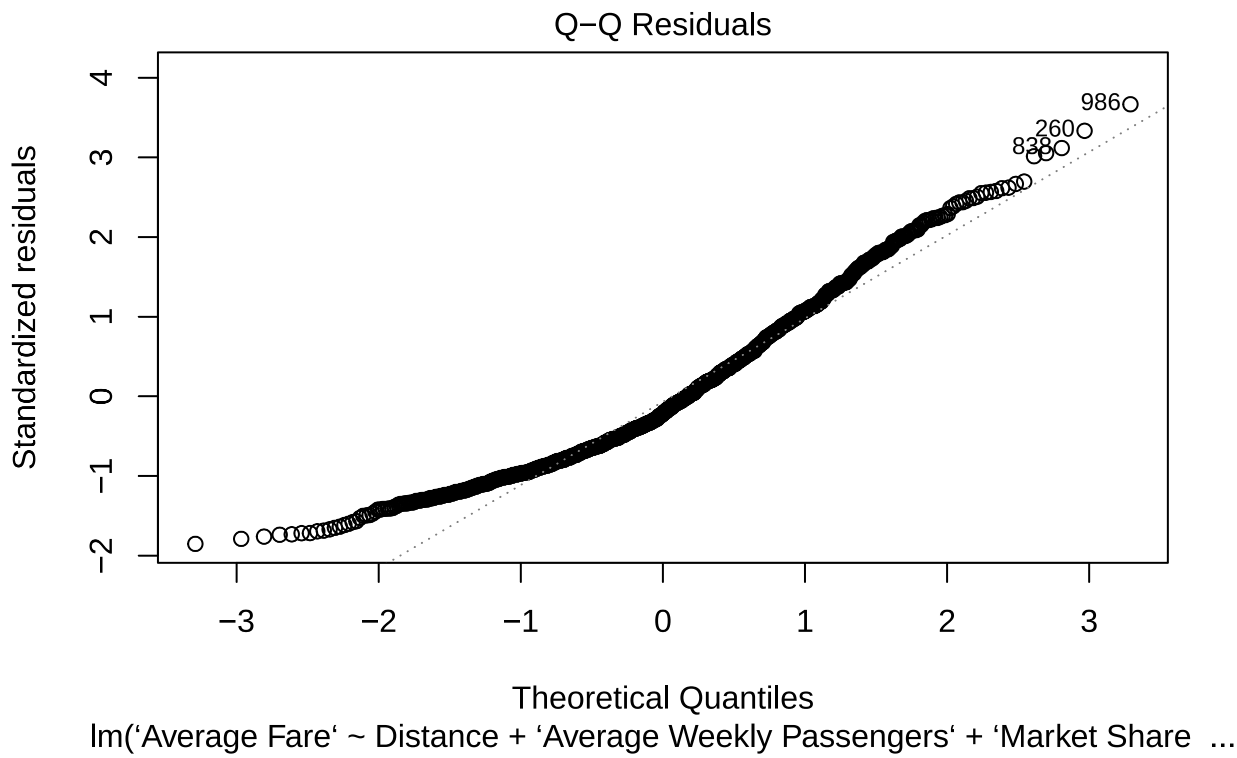
```r
summary(fit)
```

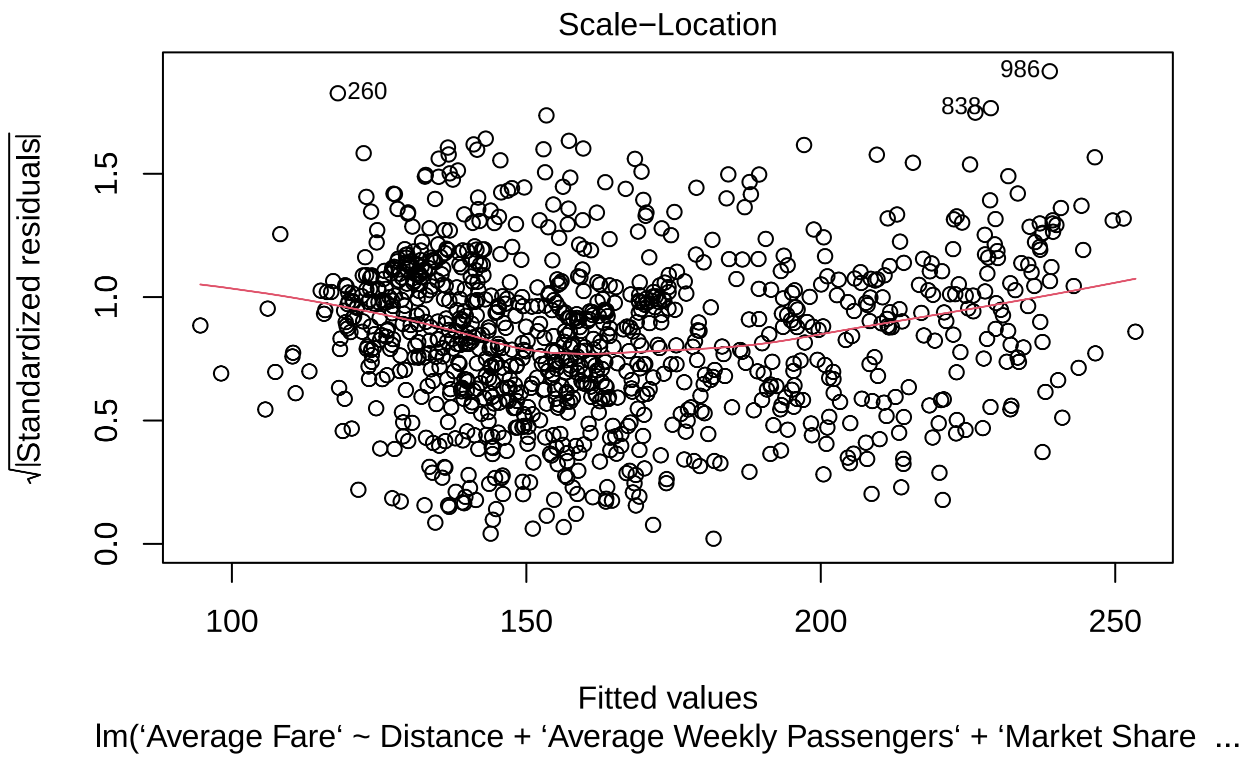
```
##
## Call:
## lm(formula = ‘Average Fare‘ ~ Distance + ‘Average Weekly Passengers‘ +
##     ‘Market Share MLA‘, data = airfares)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.108 -34.248  -9.785  28.304 162.343
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 91.803616   7.522929  12.203  < 2e-16 ***
## Distance                     0.054563   0.002611  20.898  < 2e-16 ***
## ‘Average Weekly Passengers‘ -0.004506   0.001860  -2.422  0.01560 *
## ‘Market Share MLA‘           0.281549   0.086556   3.253  0.00118 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 44.46 on 996 degrees of freedom
## Multiple R-squared:  0.357,  Adjusted R-squared:  0.3551
## F-statistic: 184.3 on 3 and 996 DF,  p-value: < 2.2e-16
```

$$AverageFare = 0.09822 \cdot Distance + 0.1929 \cdot AverageWeeklyPassengers + 1.21379 \cdot MarketShareMLA - 28.36317$$

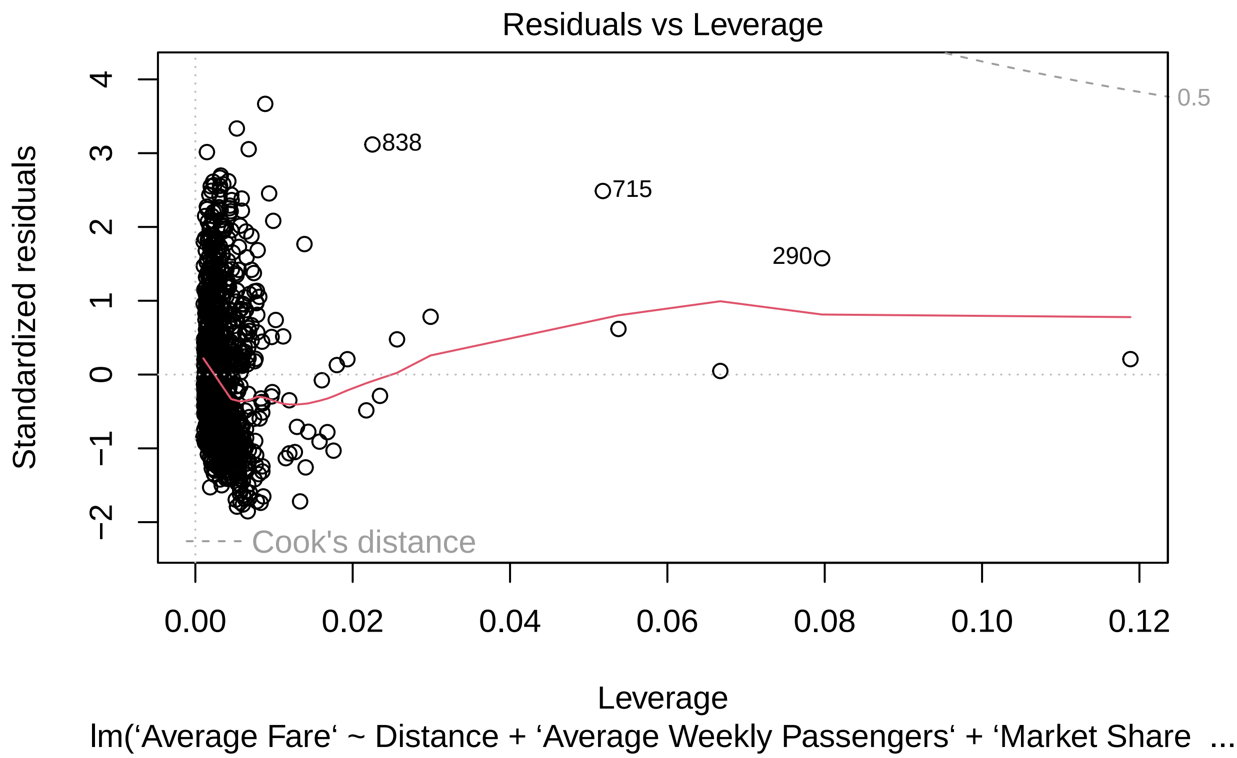
```r
plot(fit)
```


Residuals vs Fitted
Residuals
150
100
50
0
−100
100
150
200
250
Fitted values
260
986
838
lm(‘Average Fare‘ ~ Distance + ‘Average Weekly Passengers‘ + ‘Market Share ...


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(‘Average Fare‘ ~ Distance + ‘Average Weekly Passengers‘ + ‘Market Share ...
986
260
838
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3



Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
100
150
200
250
Fitted values
lm(‘Average Fare‘ ~ Distance + ‘Average Weekly Passengers‘ + ‘Market Share ...
260
838
986



2nd model

```
fit <- lm(`Average Fare` ~ Distance + `Market Share MLA`, data = airfares)
```

```
summary(fit)
```

```
##
## Call:
## lm(formula = ‘Average Fare‘ ~ Distance + ‘Market Share MLA‘,
##     data = airfares)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.745 -34.227  -9.745  28.385 159.636
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        86.051431   7.155691   12.03  < 2e-16 ***
## Distance            0.055507   0.002588   21.45  < 2e-16 ***
## ‘Market Share MLA‘  0.310252   0.085950    3.61 0.000322 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 44.57 on 997 degrees of freedom
## Multiple R-squared:  0.3532, Adjusted R-squared:  0.3519
## F-statistic: 272.2 on 2 and 997 DF,  p-value: < 2.2e-16
```

$$AverageFare = 0.09643 \cdot Distance + 1.22381 \cdot MarketShareMLA - 10.77440$$

Approach 1:

```
-28.36317 + 0.09822*1200 + 0.01929*100 + 1.21379 *100
```

```
## [1] 212.8088
```

Approach 2:

```
-10.77440 + 0.09643 * 1200 + 1.22381 * 100
```

```
## [1] 227.3226
```

Justification for choosing a certain model - either R^2 or p-value. If R^2 then first model because R^2 is higher, but if p-value number of passenger should not be included so 2nd model.