2022 02 23 VB-STA5 Reexam Solution Guide

1. Mind the gap.

a) Join the two datasets.

```
L_transport <- readr::read_csv('data/London_transport_passengers.csv')
L_ID <- readr::read_csv('data/London_transport_codes.csv')
transport <- L_transport %>% left_join(L_ID, by = c('Transportation_ID'='Transportation code'))
```

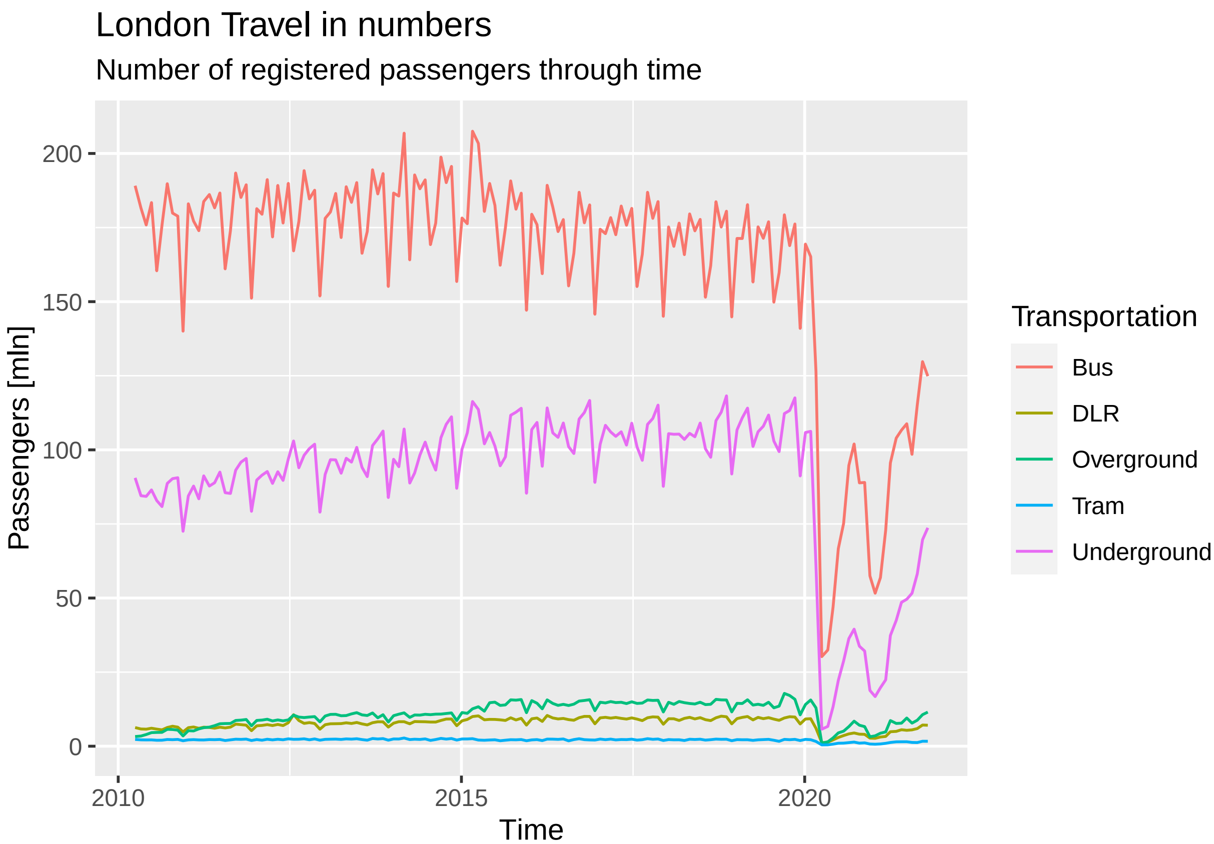
b) Present average number of passengers on all modes of transport in Reporting period *11* through the years in descending order.

```
transport %>%
  filter(`Reporting Period` == 11) %>%
  group_by(`Transportation`) %>%
  summarise(`Mean number of passengers [mln]` = mean(`Passengers [mln]`)) %>%
  arrange(desc(`Mean number of passengers [mln]`)) %>%
  knitr::kable()
```

Transportation	Mean number of passengers [mln]
Bus	166.267828
Underground	91.501390
Overground	11.170393
DLR	7.818876
Tram	2.142591

c) Recreate the plot.

```
transport %>%
  ggplot() +
  geom_line(aes(x = `Period beginning`, y = `Passengers [mln]`, color = Transportation)) +
  labs(title = 'London Travel in numbers',
       subtitle = 'Number of registered passengers through time',
       x = 'Time')
```



d) Describe the plot.

- presents number of registered passenger in different modes of transport in London in years 2010 - 2021
- data points are connected lineary
- there are visible variations in number of passengers through time.
- most passengers travel on bus, then underground. remaining 3 take much smaller part of share
- there is vissible drop in travels at the beginning of 2020 - most likely due to the global pandemic
- the transport is slowly recovering, but the numbers are nowhere near the numbers before pandemic

e) In 2017 a group of students at Imperial College London conducted a survey of 927 London commuters. The survey asked questions about service satisfaction using a couple of carefully formed questions. The students asked random travelers to answer the questions. Here is a summary of how many people were surveyed in the different modes of transport:

```
transport_2017 <- transport %>% filter(Year == 2017)
all_passengers <- sum(transport_2017$`Passengers [mln]`)
(proportions <- transport_2017 %>%
  group_by(Transportation) %>%
  summarize(sum = sum(`Passengers [mln]`)) %>%
  mutate(proportion = sum/all_passengers))
```

```
## # A tibble: 5 x 3
##   Transportation    sum proportion
##   <chr>           <dbl>      <dbl>
## 1 Bus             2253.      0.570
```

```
## 2 DLR              121.     0.0307
## 3 Overground       191.     0.0482
## 4 Tram              29.5    0.00746
## 5 Underground    1357.     0.343
```

```
(students <- tribble(
  ~`Mode of Transport`, ~`No. of passengers interviewed`,
  "Bus",   461,
  "Underground",   347,
  "DLR",   26,
  "Tram",   25,
  "Overground", 68
) )
```

```
## # A tibble: 5 x 2
##   `Mode of Transport` `No. of passengers interviewed`
##   <chr>                                         <dbl>
## 1 Bus                                             461
## 2 Underground                                     347
## 3 DLR                                              26
## 4 Tram                                             25
## 5 Overground                                       68
```

Chi square test for goodness of fit.

H0: Distribution of passengers intervied within different modes of transport is the same as distribution of whole passengers in 2017.

HA: Distribution of passengers intervied within different modes of transport is not the same as distribution of whole passengers in 2017.

alpha significance level - 0.05

Conditions check:

- we assume that the dataset is independent
- expected cases should be more than 5

```
students <- students %>%  left_join(proportions, by=c('Mode of Transport'='Transportation'))

(students <- students %>% mutate(expected = proportion*927))
```

```
## # A tibble: 5 x 5
##   `Mode of Transport` `No. of passengers interviewed`    sum proportion expected
##   <chr>                                         <dbl>  <dbl>      <dbl>    <dbl>
## 1 Bus                                             461 2253.     0.570     529.
## 2 Underground                                     347 1357.     0.343     318.
## 3 DLR                                              26  121.     0.0307     28.4
## 4 Tram                                             25   29.5    0.00746     6.92
## 5 Overground                                       68  191.     0.0482     44.7
```

All expected values are above 5.

- short version

```r
chisq.test(students$`No. of passengers interviewed`, p=students$proportion)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  students$`No. of passengers interviewed`
## X-squared = 70.795, df = 4, p-value = 1.542e-14
```

We reject null hypothesis in favour of alternative. Distribution of passengers intervied within different modes of transport is not the same as distribution of whole passengers in 2017.

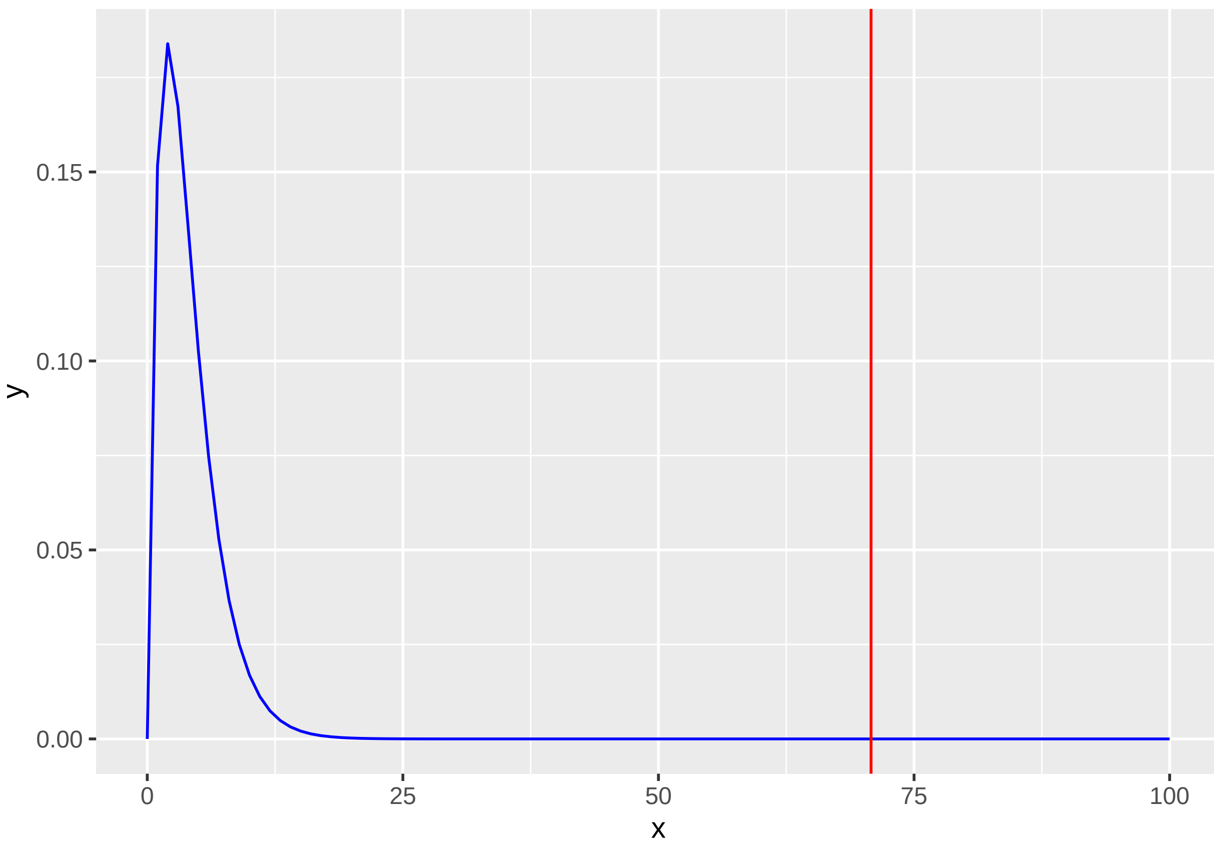
- long version

```r
(chi2_stat <- sum(((students$`No. of passengers interviewed` - students$expected)^2)/students$expected))
```

```
## [1] 70.79514
```

```r
dof <- 4
```

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



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

```
## [1] 1.54321e-14
```

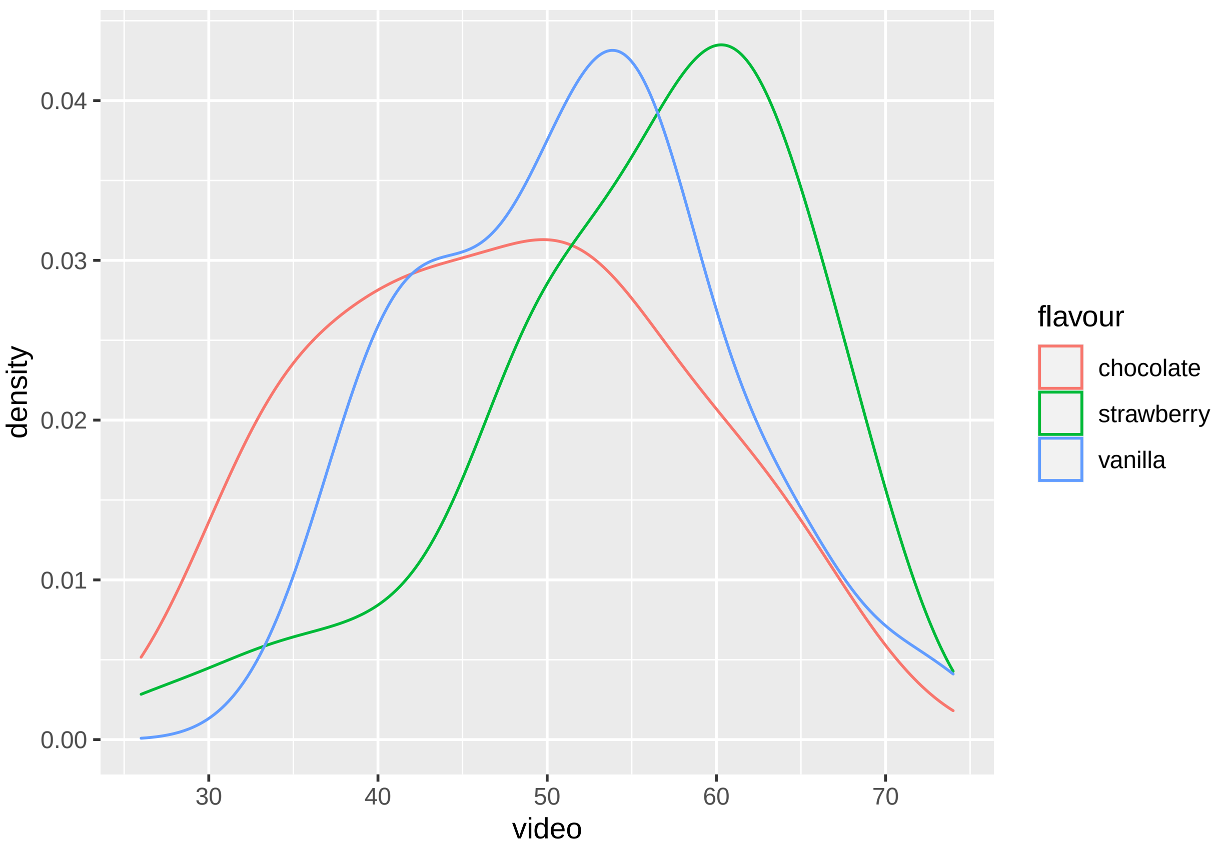
We reject null hypothesis in favour of alternative. Distribution of passengers intervied within different modes of transport is not the same as distribution of whole passengers in 2017.

2. Ice cream

Dataset *data/ice_cream.csv* contains information from an experiment, where a group of people were asked to choose in between three flavours of ice cream - strawberry, chocolate, and vanilla. Subsequently, they have been evaluated in playing video games and doing puzzles.

a) Present distribution function of video games scores divided according to ice cream flavour.

```r
ice_cream <- readr::read_csv('data/ice_cream.csv')
ice_cream %>%
  ggplot() +
  geom_density(aes(x = video, color = flavour))
```



b) Is there a statistically significant difference between the mean puzzle score for males with vanilla preference vs. males with strawberry preference? Conduct a suitable test.

Difference of means t-test.

$H_0 : \mu_{male_vanilla} - \mu_{male_strawberry} = 0$

$H_A : \mu_{male_vanilla} - \mu_{male_strawberry} = 0 \neq 0$

H0: There is no difference between mean puzzle score for male prefering vanilla and strawberry flavours.

HA: There is a difference between mean puzzle score for male prefering vanilla and strawberry flavours.

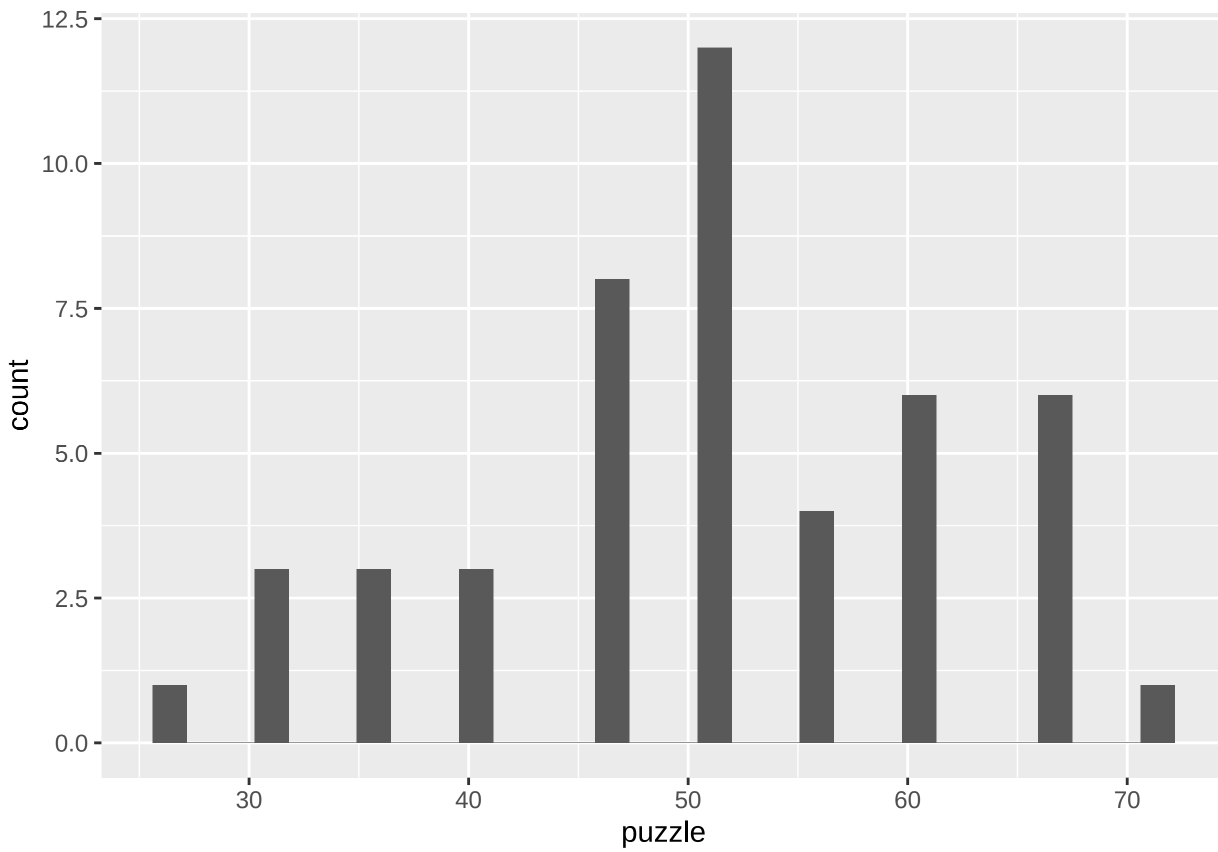
alpha significance level - 0.05

Conditions check:

Normality:

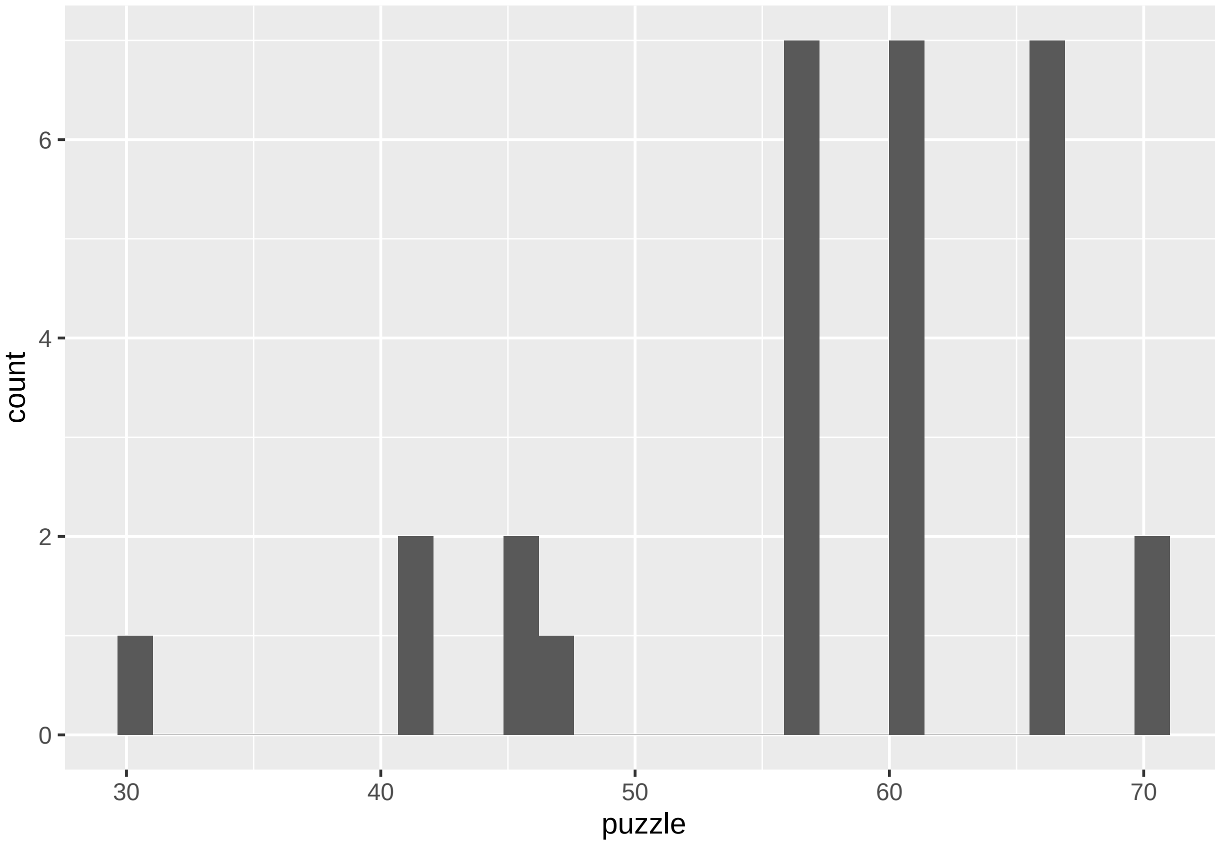
```
ice_cream %>%
  filter(flavour == 'vanilla') %>%
  filter(sex == 'M') %>%
  ggplot() +
  geom_histogram(aes(x = puzzle))
```

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




```
ice_cream %>%
  filter(flavour == 'strawberry') %>%
  filter(sex == 'M') %>%
  ggplot() +
  geom_histogram(aes(x = puzzle))
```

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



Hard to say with samples so small (strawberry), but it seems that most variables follow normal distribution.

We assume that observations are independent.

- short version

```
ice_cream %>%
  filter(sex == 'M', flavour %in% c('vanilla', 'strawberry')) %>%
  t.test(puzzle~flavour, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  puzzle by flavour
## t = 2.9419, df = 65.277, p-value = 0.00451
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   2.250291 11.761448
## sample estimates:
## mean in group strawberry    mean in group vanilla
##                 57.79310                 50.78723
```

p-value is smaller than alpha significance level, thus we reject null hypothesis and accept the alternative. There is statistically significant difference between mean puzzle score for male prefering vanilla and strawberry flavours.

- long version

```
m_v <- ice_cream %>%
  filter(flavour == 'vanilla') %>%
  filter(sex == 'M')
m_s <- ice_cream %>%
  filter(flavour == 'strawberry') %>%
  filter(sex == 'M')

(point_estimate <- mean(m_v$puzzle) - mean(m_s$puzzle))
```

```
## [1] -7.005869
```

```
(nrow(m_v))
```

```
## [1] 47
```

```
(nrow(m_s))
```

```
## [1] 29
```

```
dof <- 28
```

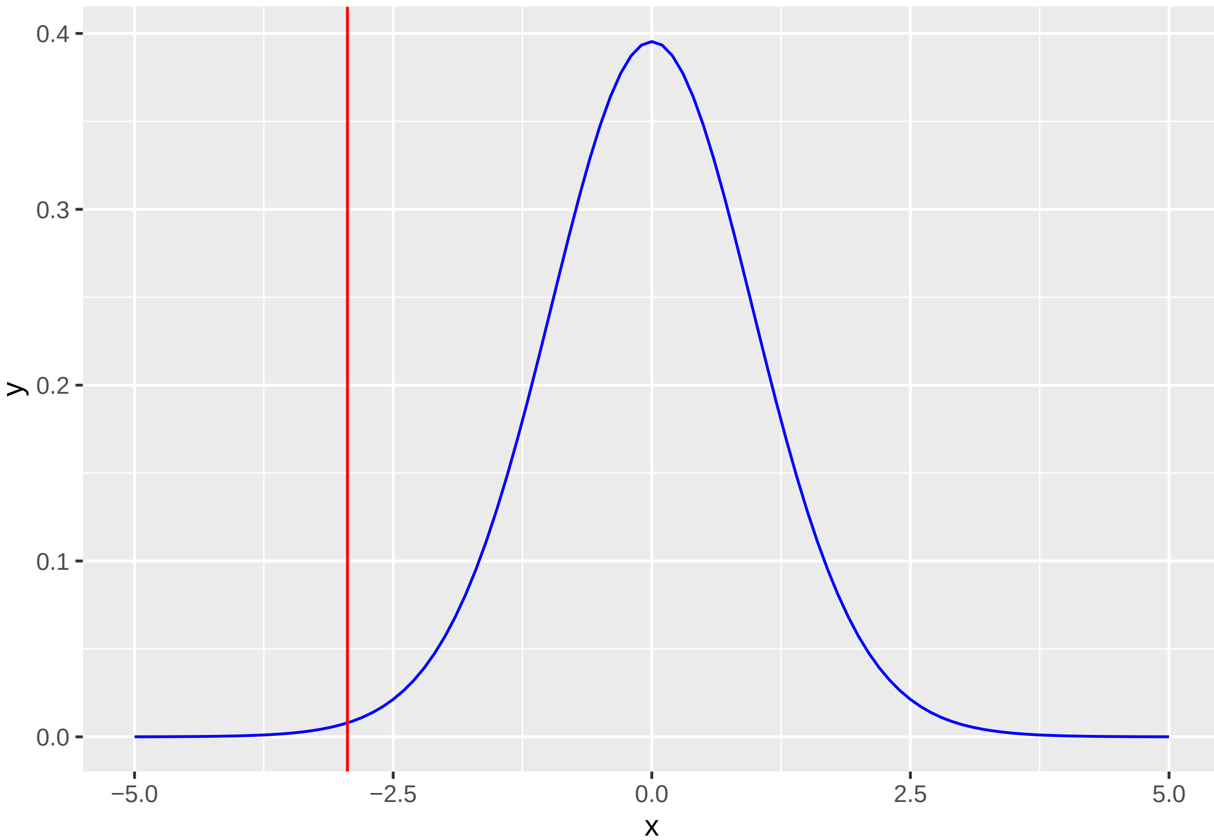
```
(SE <- sqrt((sd(m_v$puzzle)^2/nrow(m_v)) + (sd(m_s$puzzle)^2/nrow(m_s))))
```

```
## [1] 2.381388
```

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

```
## [1] -2.941926
```

```
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 * pt(t_score, df = dof))
```

```
## [1] 0.006481958
```

p-value is smaller than alpha significance level, thus we reject null hypothesis and accept the alternative. There is statistically significant difference between mean puzzle score for male prefering vanilla and strawberry flavours.

3. Health insurance

Dataset *data/insurance.csv* contains information about over 1000 randomly chosen U.S. participants. Their insurance packages range in between low-cost insurance - up to \$15.000 per year, medium-cost insurance \$15.000 - \$30.000 per year, and high-cost insurance - above \$30.000 per year.

a) What are statistically significant predictors of the high-cost insurance? Create a model and tune it.

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

```
## Rows: 1338 Columns: 7
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): sex, smoker, region
## dbl (4): age, bmi, children, charges
```

```
##
## 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.
```

```
insurance_high <- insurance %>%  filter(charges>30000)
colnames(insurance_high)
```

```
## [1] "age"      "sex"      "bmi"      "children" "smoker"   "region"   "charges"
```

To tune the model to get the most statistically significant model we use p-value approach. In this case backwareds elimination.

```
fit <- lm(charges ~ age +
            sex +
            bmi +
            children +
            region +
            smoker
          , data = insurance_high)
summary(fit)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + region +
##     smoker, data = insurance_high)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12783.1  -1438.8   -585.1    374.3  22937.0
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       850.22    2948.61   0.288    0.773
## age               258.77      22.50  11.499  < 2e-16 ***
## sexmale          -710.87     648.01  -1.097    0.274
## bmi               550.15      71.35   7.711 1.48e-12 ***
## children         -179.93     283.41  -0.635    0.526
## regionnorthwest  1365.10    1006.20   1.357    0.177
## regionsoutheast  -256.66     868.82  -0.295    0.768
## regionsouthwest   377.07     929.17   0.406    0.685
## smokeryes       11426.06    1376.03   8.304 4.98e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3924 on 153 degrees of freedom
## Multiple R-squared:  0.6152, Adjusted R-squared:  0.5951
## F-statistic: 30.57 on 8 and 153 DF,  p-value: < 2.2e-16
```

Remove region

```r
fit <- lm(charges ~ age +
            sex +
            bmi +
            children +
            #region +
            smoker
          , data = insurance_high)
summary(fit)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker, data = insurance_high)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13636  -1296   -660    240  22459
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2538.41    2773.72   0.915    0.362
## age           257.04      22.46  11.443  < 2e-16 ***
## sexmale      -827.50     644.43  -1.284    0.201
## bmi           515.29      66.62   7.735 1.20e-12 ***
## children      -89.56     278.08  -0.322    0.748
## smokeryes   11211.71    1368.69   8.192 8.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3925 on 156 degrees of freedom
## Multiple R-squared:  0.6075, Adjusted R-squared:  0.5949
## F-statistic: 48.28 on 5 and 156 DF,  p-value: < 2.2e-16
```

Remove children.

```r
fit <- lm(charges ~ age +
            sex +
            bmi +
            #children +
            #region +
            smoker
          , data = insurance_high)
summary(fit)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + smoker, data = insurance_high)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13800.1  -1269.1   -603.0    214.2  22562.4
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2485.25    2760.89   0.900    0.369
## age           256.11      22.21  11.531  < 2e-16 ***
## sexmale      -840.08     641.40  -1.310    0.192
## bmi           513.72      66.25   7.754 1.05e-12 ***
## smokeryes   11263.77    1355.23   8.311 4.20e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3913 on 157 degrees of freedom
## Multiple R-squared:  0.6072, Adjusted R-squared:  0.5972
## F-statistic: 60.67 on 4 and 157 DF,  p-value: < 2.2e-16
```

Remove sex.

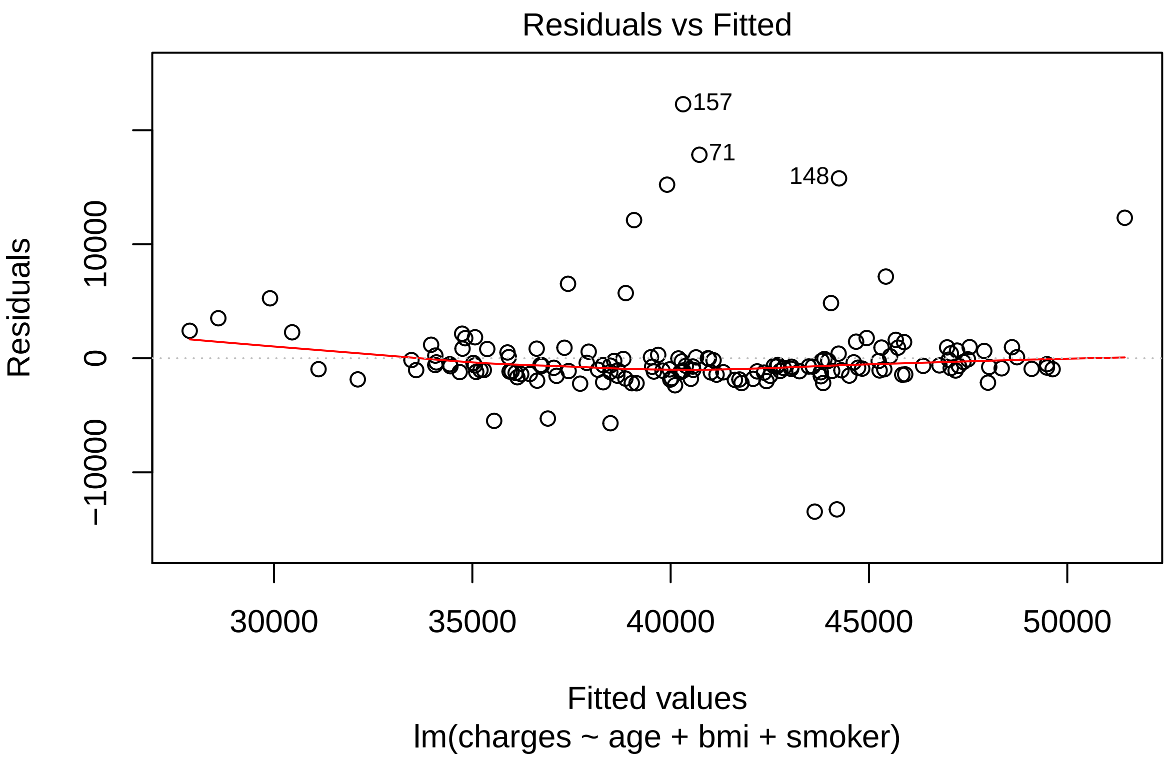
```
fit <- lm(charges ~ age +
            #sex +
            bmi +
            #children +
            #region +
            smoker
          , data = insurance_high)
summary(fit)
```

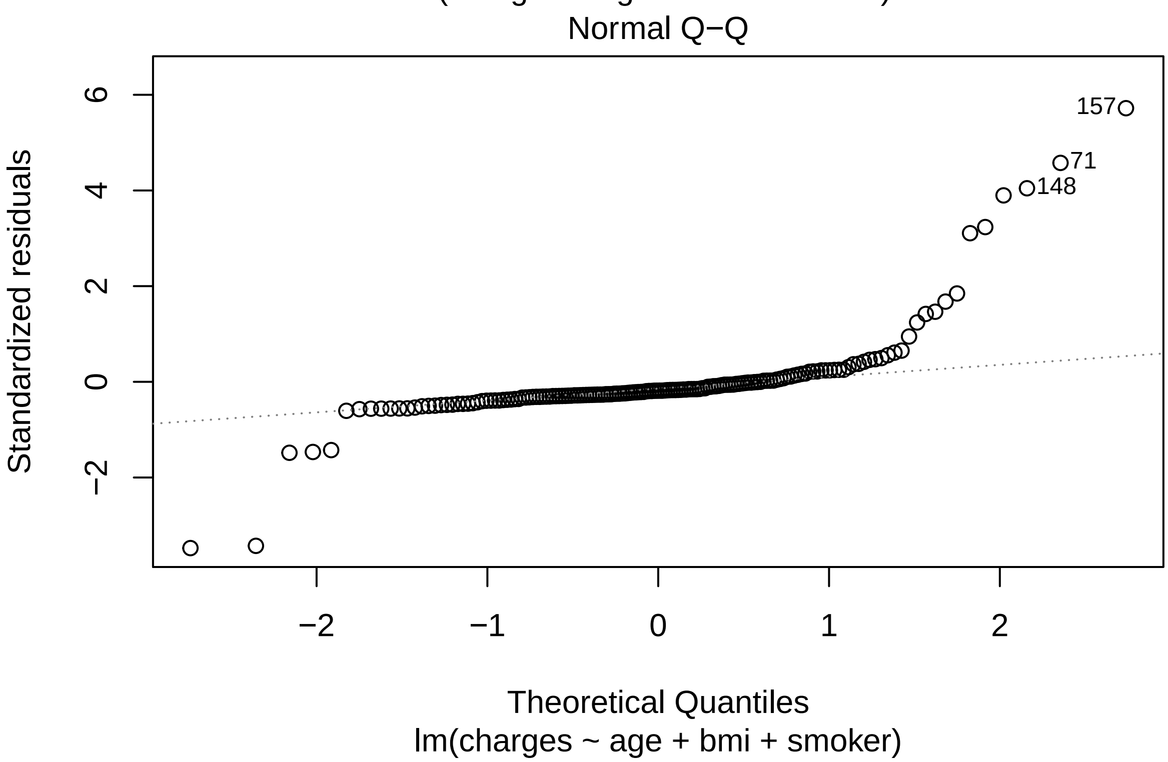
```
##
## Call:
## lm(formula = charges ~ age + bmi + smoker, data = insurance_high)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13447.1  -1205.8   -723.4     99.0  22279.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1910.63    2731.98   0.699    0.485
## age           256.64      22.26  11.531  < 2e-16 ***
## bmi           517.67      66.33   7.804 7.70e-13 ***
## smokeryes   11137.27    1354.84   8.220 6.94e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3922 on 158 degrees of freedom
## Multiple R-squared:  0.6029, Adjusted R-squared:  0.5954
## F-statistic: 79.96 on 3 and 158 DF,  p-value: < 2.2e-16
```

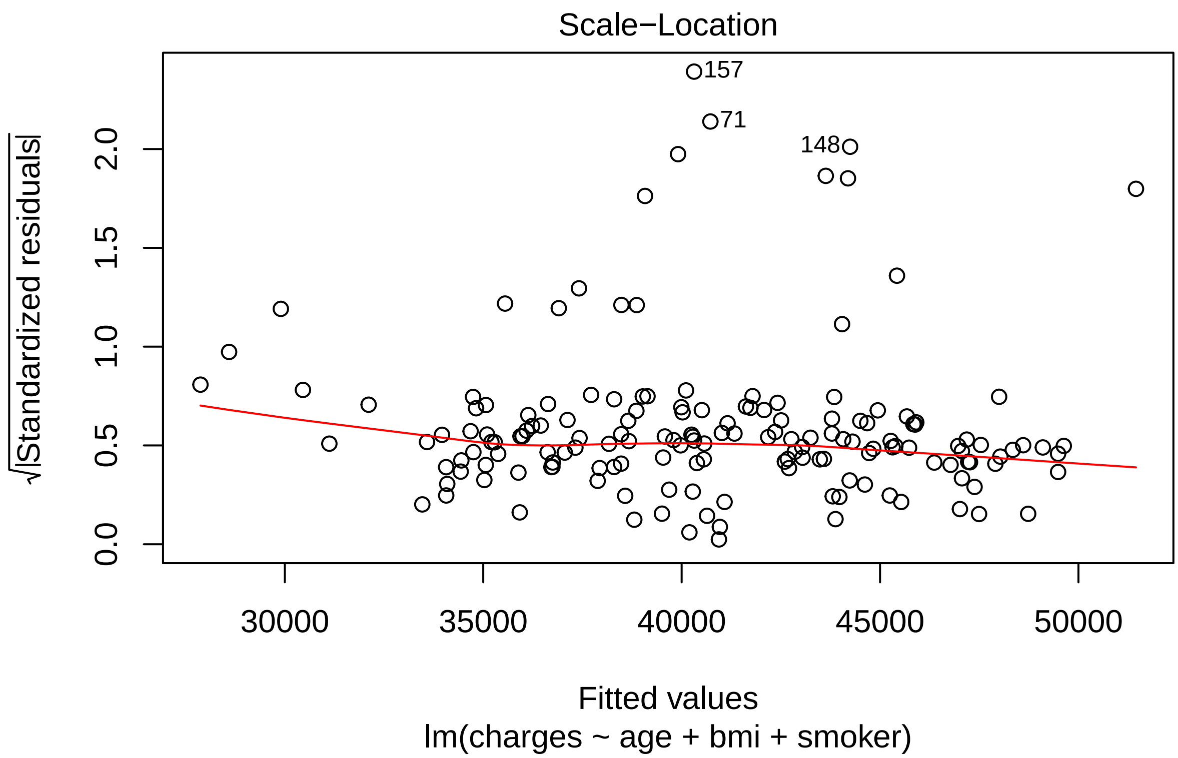
b) Evaluate the model.

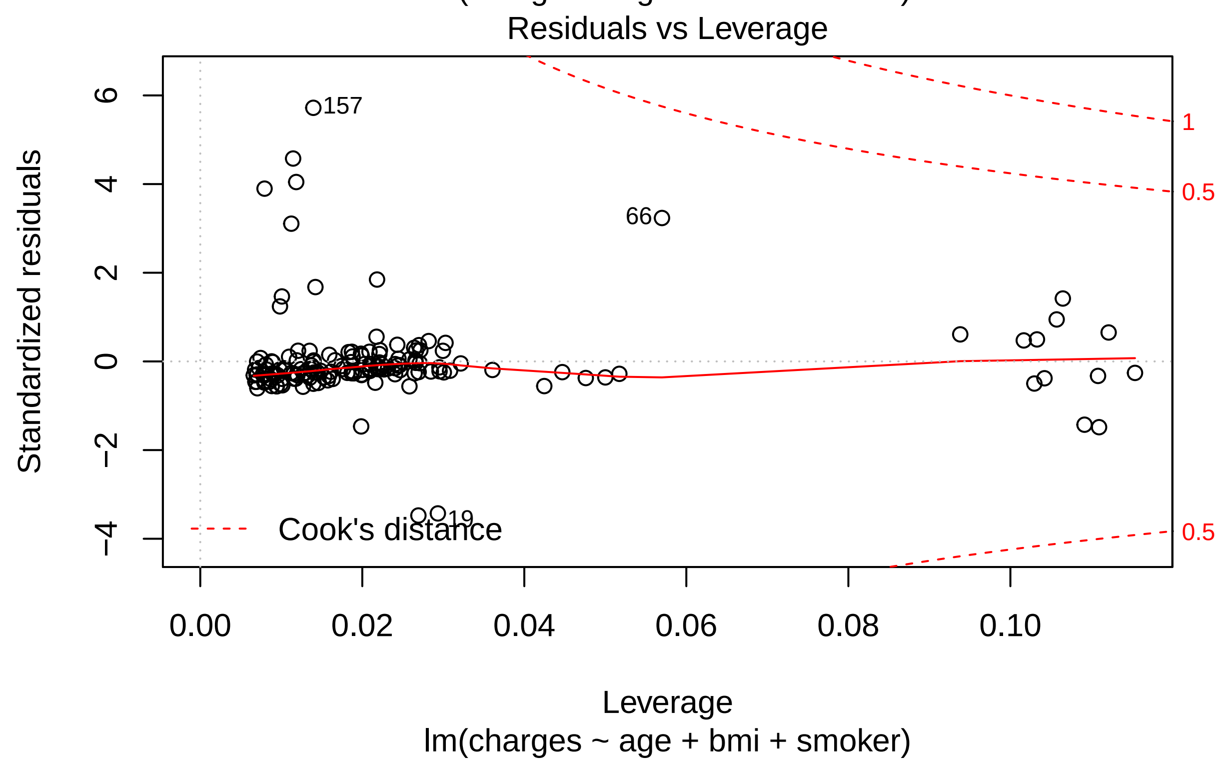
- constant residuals and normal distribution of residuals

```
plot(fit)
```


Residuals vs Fitted
157
71
148
Residuals
10000
0
−10000
30000
35000
40000
45000
50000
Fitted values
lm(charges ~ age + bmi + smoker)


Normal Q−Q
157
71
148
Standardized residuals
6
4
2
0
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(charges ~ age + bmi + smoker)

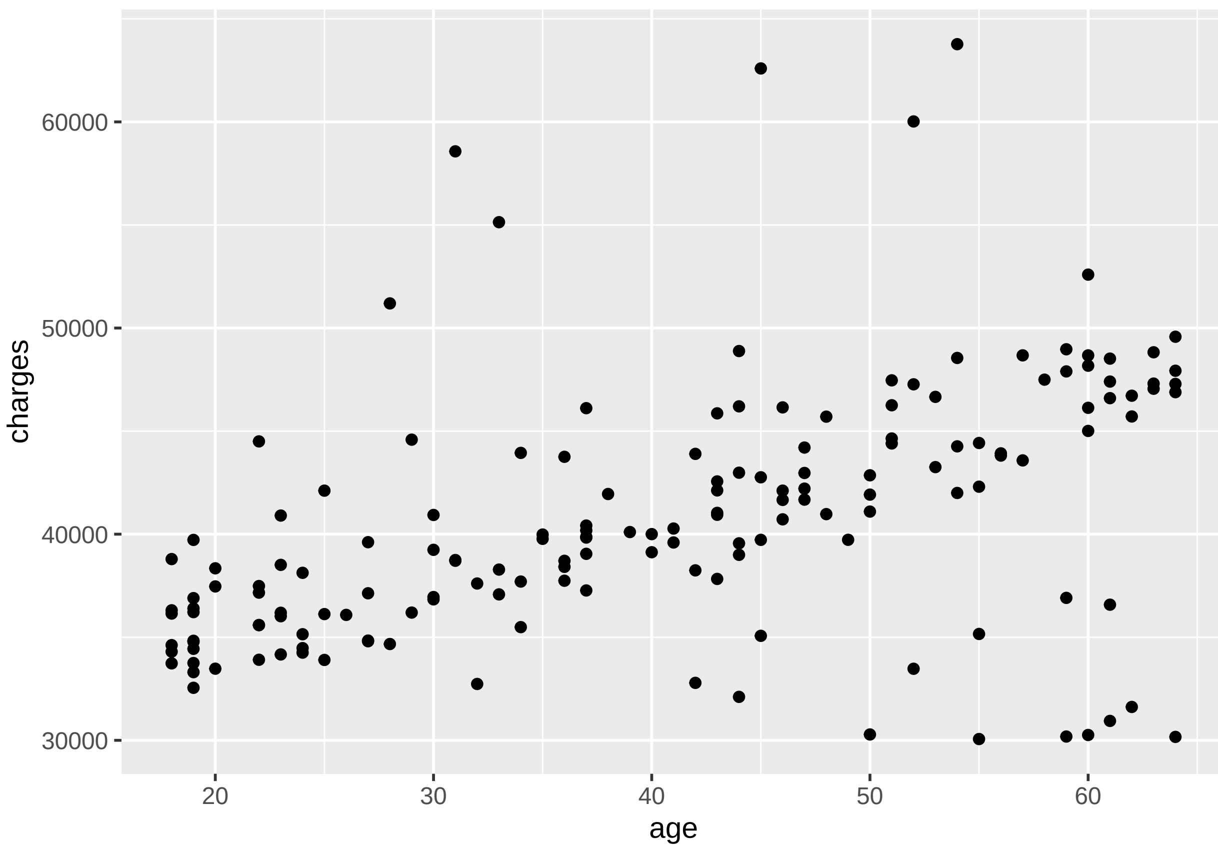




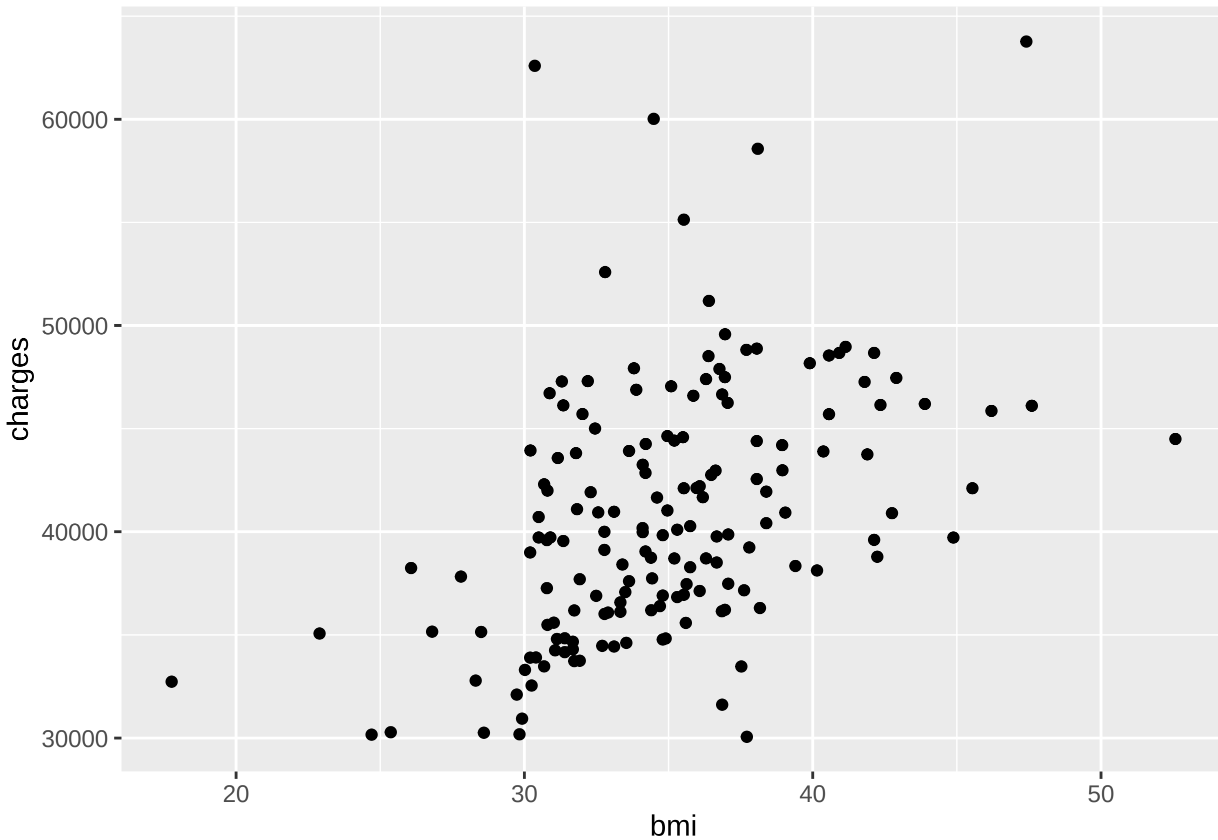
As seen at the first two plots, there are some outliers from normal distribution of the residuals, however majority of points follow theoretical line. The residuals also seem to be evanly distributed.

- we assume independence
- linearity

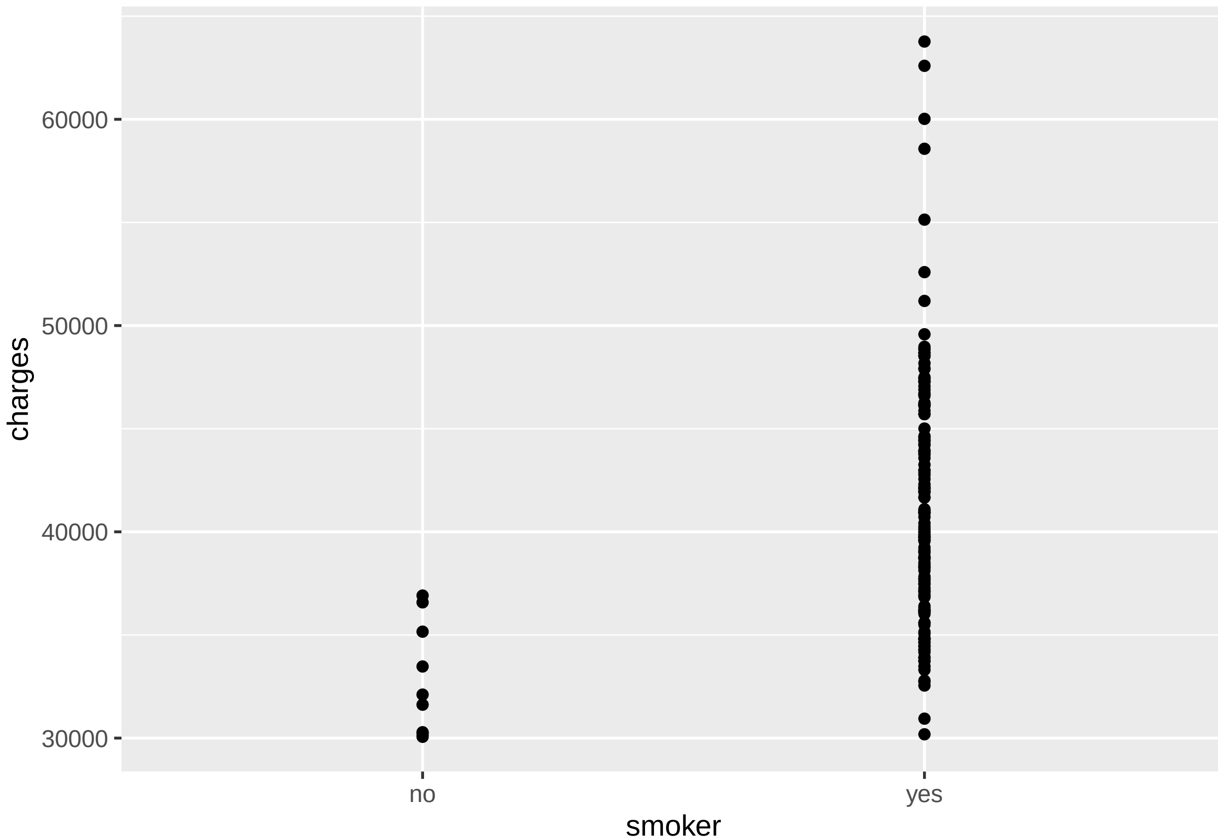
```r
ggplot(insurance_high) +
  geom_point(aes(x = age, y = charges))
```




```r
ggplot(insurance_high) +
  geom_point(aes(x = bmi, y = charges))
```




```r
ggplot(insurance_high) +
  geom_point(aes(x = smoker, y = charges))
```



There seems to be linear corelation visible for age and smoker, slightly vissible for bmi.