2023 02 22 VB-STA5 Exam in Statistics

Wednesday 22nd of February.

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. Gymnastics and figure skating.

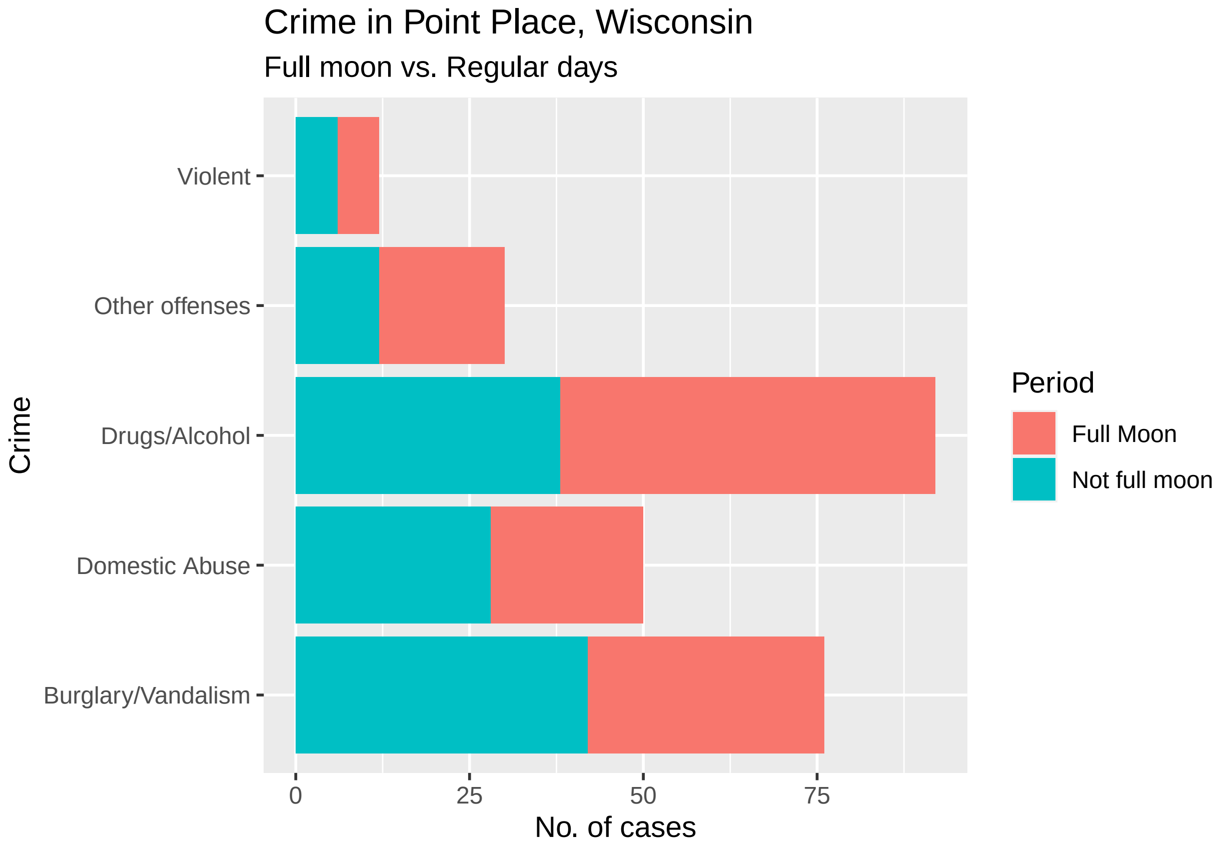
Dataset *data/gym_figskate.csv* contains information about Olympic athletes in Figure Skating and Gymnastics in years 1964 to 2016.

a) Recreate the plot.

```
moon <- readr::read_csv("data/full_moon.csv")
```

```
## Rows: 260 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Offense, Period
##
## 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.
```

```
moon %>%
  ggplot() +
  geom_bar(aes(y = Offense, fill = Period))  +
  labs(subtitle = 'Full moon vs. Regular days ',
       title = 'Crime in Point Place, Wisconsin',
       x = 'No. of cases',
       y = 'Crime',
       fill = 'Period')
```



b) Describe the plot, including (but not limited to) comparison of the crime types according to moon phase.

- Bar plot showing number of particular crimes comited in Point Place Wisconsin.
- Colored according to time peroiod - Full Moon and Not full moon days.
- The most typical crime in Point Place is Drugs/Alcohol related. Then Burglary/Vandalism, Domestic Abuse and the rarest are Violent crimes.
- It seems that there are more Drug/Alcohol offences happening during the Full Moon phase, but the difference is not very big.
- Domestic Abuse, Burglary/Vandalism has opposite tendency - more on regular days as opposed to the full moon.
- Violent Crimes are spread evenly.

c) Is there statistical relation in between the crimes committed and moon phase? Which test can you use to check it? What are conditions of this test to be valid? Conduct the test and form statistical conclusions.

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 the crimes committed and moon phase.

HA: There is correlation in between the crimes committed and moon phase.

alpha significance level - 0.05

```
fm <- moon %>% filter( Period == 'Full Moon') %>% group_by(Offense) %>% tally()
colnames(fm)[2] <- 'Full moon'
nfm <- moon %>% filter( Period != 'Full Moon') %>% group_by(Offense) %>% tally()
colnames(nfm)[2] <- 'Not full moon'
m <- fm %>% left_join(nfm, by = c('Offense'))
```

```
(sum_all <- sum(m$`Full moon`) + sum(m$`Not full moon`))
```

```
## [1] 260
```

```
m <- m %>% mutate(fm_expected = (sum(m$`Full moon`)*(`Full moon`+`Not full moon`)/sum_all))
m <- m %>% mutate(nfm_expected = (sum(m$`Not full moon`)*(`Full moon`+`Not full moon`)/sum_all))
```

```
m
```

```
## # A tibble: 5 x 5
##   Offense            ‘Full moon‘ ‘Not full moon‘ fm_expected nfm_expected
##   <chr>                    <int>           <int>       <dbl>        <dbl>
## 1 Burglary/Vandalism          34              42       39.2         36.8
## 2 Domestic Abuse              22              28       25.8         24.2
## 3 Drugs/Alcohol               54              38       47.4         44.6
## 4 Other offenses              18              12       15.5         14.5
## 5 Violent                      6               6        6.18         5.82
```

All expected values are above 5.

- short version

```
m %>% select(2,3) %>%
chisq.test()
```

```
##
##  Pearson’s Chi-squared test
##
## data:  .
## X-squared = 5.3036, df = 4, p-value = 0.2575
```

We accept null hypothesis and reject the alternative. There is no correlation in between the crimes committed and moon phase.

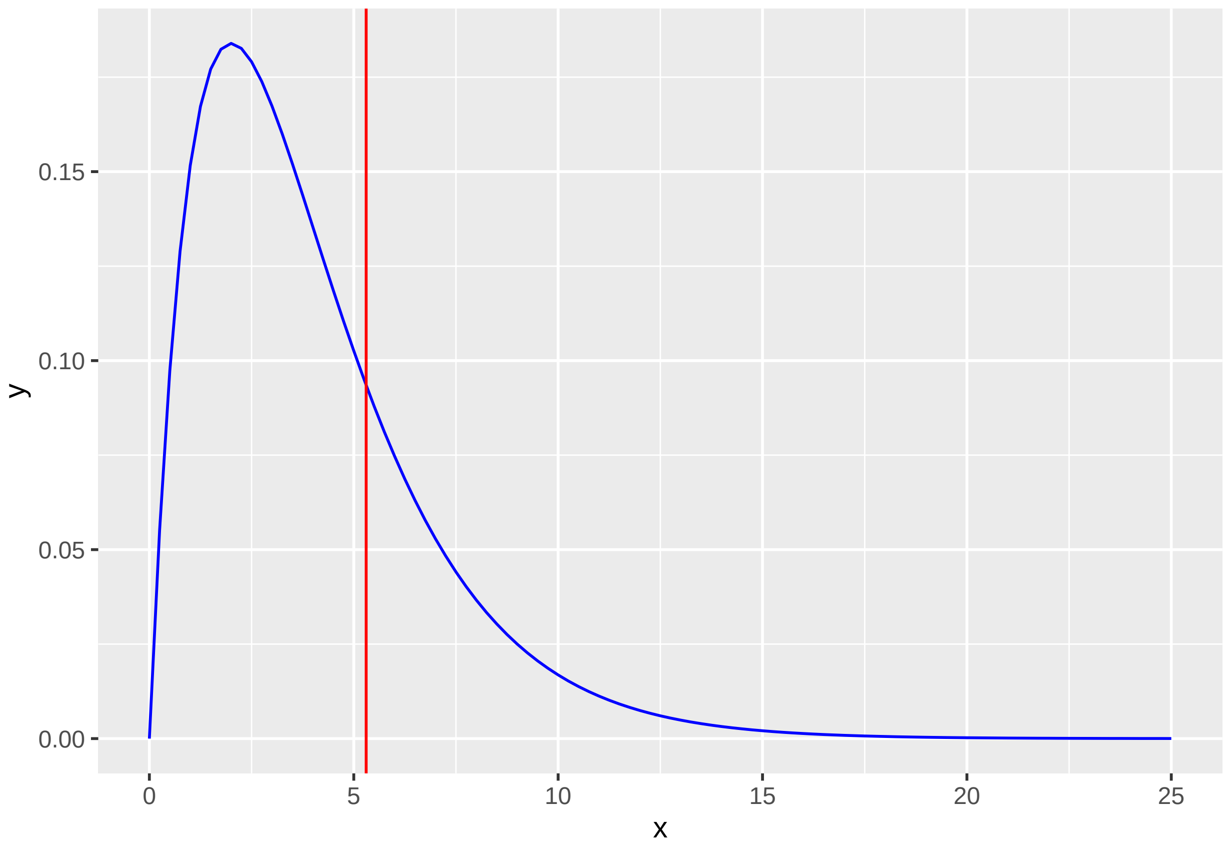
- long version

```
(chi2_stat <- sum(((m$`Full moon` - m$fm_expected)/ m$fm_expected^0.5)^2) +
             sum(((m$`Not full moon` - m$nfm_expected)/ m$nfm_expected^0.5)^2))
```

```
## [1] 5.303581
```

```
dof <- 4
```

```
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] 0.2575419
```

We accept null hypothesis and reject the alternative. There is no correlation in between the crimes committed and moon phase.

2. Salmon farming

Dataset *data/salmon.csv* contains information about dangerous chemical compounds found in salmon selected randomly from various salmon farms around the world.

a) Calculate mean amount of highly cancerogenous substances - PCBs (Total PCBs) per country of salmon origin. Present in ascending order in format presented below.

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

```
## Rows: 153 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Kind, Location
## dbl (11): Mirex, Hexachlorobenzene, HCH_gamma, Heptachlor Epoxide, Dieldrin,...
```

```
##
## 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.
```

```
salmon %>% group_by(Location) %>%
  summarize(`Mean PCBs Contents` = mean(`Total PCBs`)) %>%
  arrange(`Mean PCBs Contents`) %>%
  knitr::kable()
```

Location	Mean PCBs Contents
Washington	18277.78
Chile	18441.67
Maine	30100.00
Western Canada	33872.22
Eastern Canada	38675.00
Norway	41641.67
Faroe Islands	47883.33
Scotland	50640.00

b) Is there a statistically significant difference in between total PCBs contents in salmon from Scotland and from Norway? Which test is appropriate to use here? What are its conditions? Conduct the test and form statistical conclusions.

The difference of means t-test. The conditions are:

- samples are independent
- population is not strongly skewed

$H_0 : \mu_{m\_Scottland} - \mu_{m\_Norway} = 0$

$H_A : \mu_{m\_Scottland} - \mu_{m\_Norway} = 0 \neq 0$

H0: There is no difference between mean PCBs content in salmon from Scotland and Norway.

HA: There is a difference between mean height for male figure scaters in Sapporo and Salt Lake City Olympics.

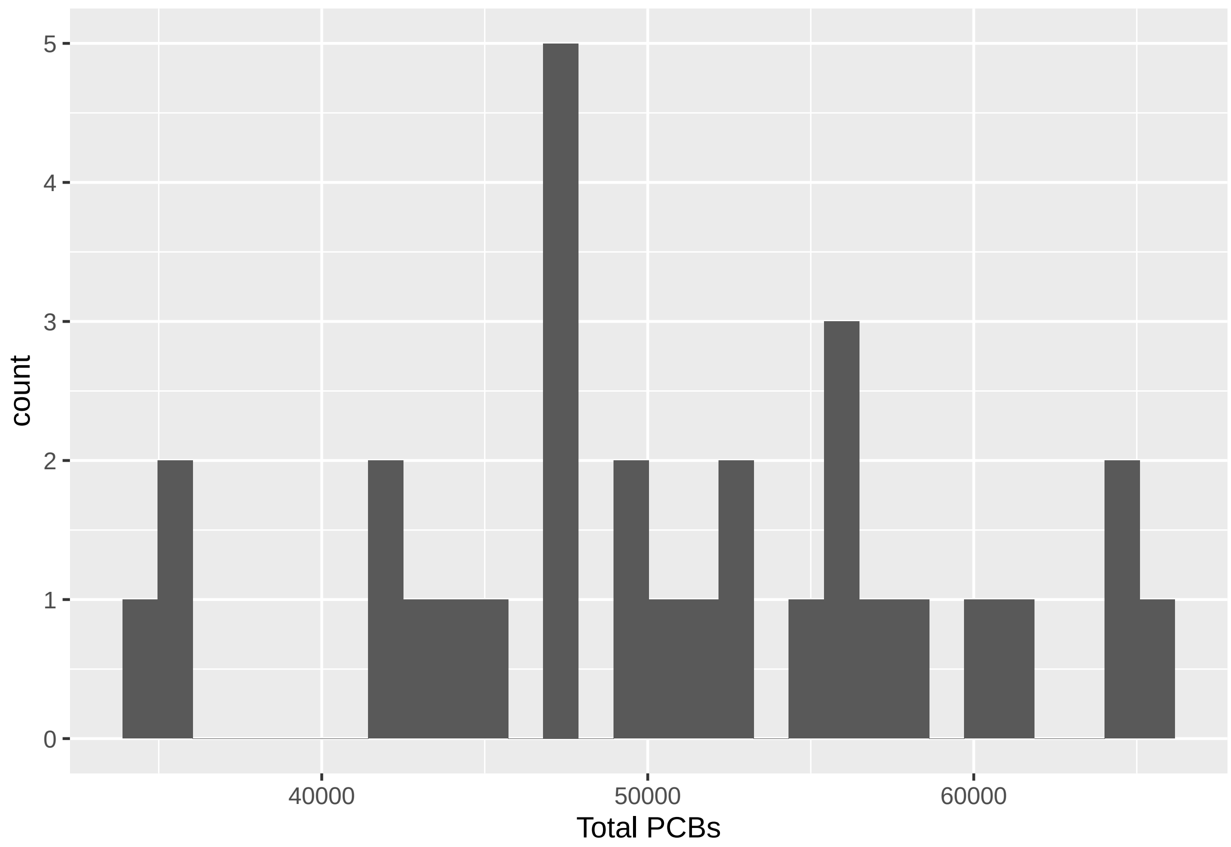
alpha significance level - 0.05

Conditions check:

Normality:

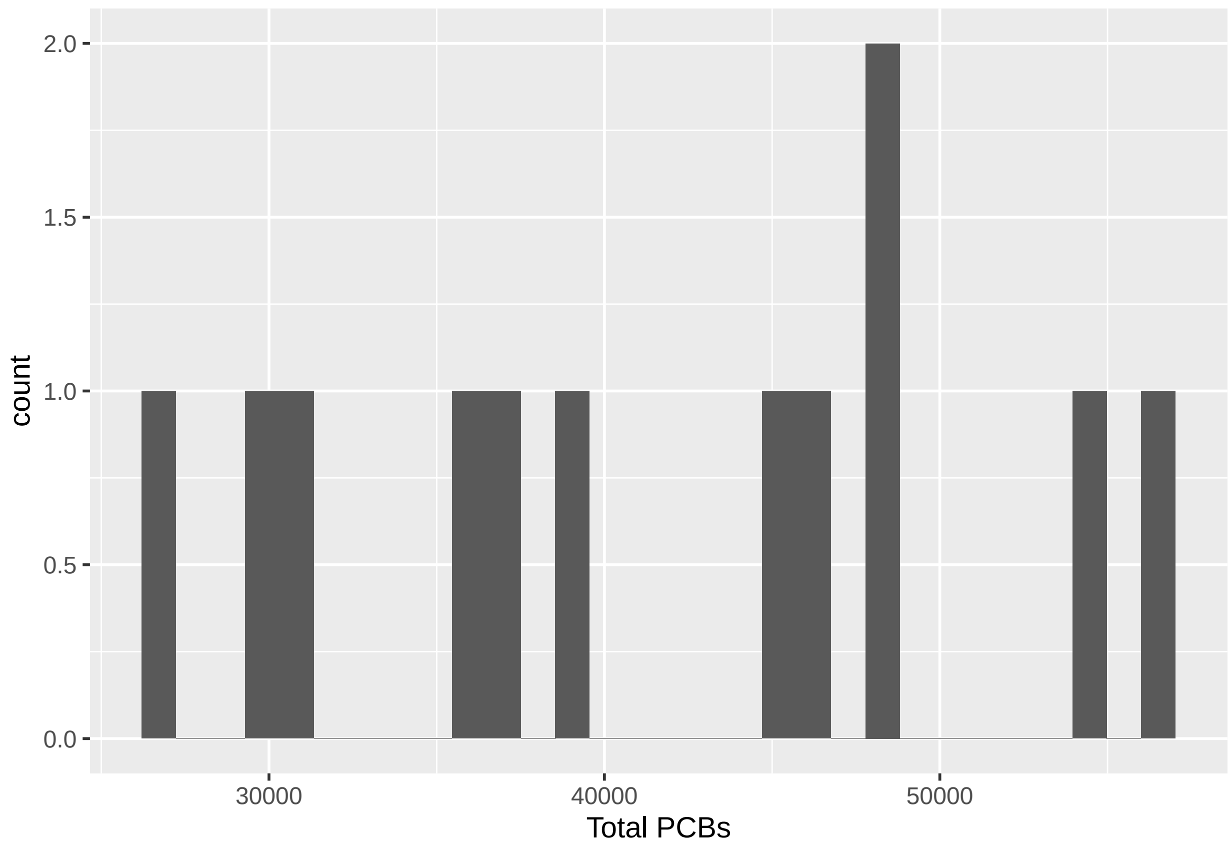
```
salmon %>% filter(Location == 'Scotland') %>%
  ggplot() +
  geom_histogram(aes(x = `Total PCBs`))
```

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



```
salmon %>% filter(Location == 'Norway') %>%
  ggplot() +
  geom_histogram(aes(x = `Total PCBs`))
```

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



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

We assume that observations are independent.

- short version

```
salmon %>%
  filter(Location %in% c('Scotland', 'Norway')) %>%
  t.test(`Total PCBs`~Location, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  Total PCBs by Location
## t = -2.7765, df = 17.778, p-value = 0.01255
## alternative hypothesis: true difference in means between group Norway and group Scotland is not equal
## 95 percent confidence interval:
##  -15813.419  -2183.248
## sample estimates:
##   mean in group Norway mean in group Scotland
##               41641.67               50640.00
```

p-value is smaller than alpha significance level, thus we reject null hypothesis in favor of the alternative. There is statistically significant difference between mean PCBs in salmon from Scotland and Norway.

- long version

```
s_s <- salmon %>% filter(Location == 'Scotland')
s_n <- salmon %>% filter(Location == 'Norway')

(point_estimate <- mean(s_s$`Total PCBs`) - mean(s_n$`Total PCBs`))
```

```
## [1] 8998.333
```

```
(nrow(s_s))
```

```
## [1] 30
```

```
(nrow(s_n))
```

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

```
dof <- 11
```

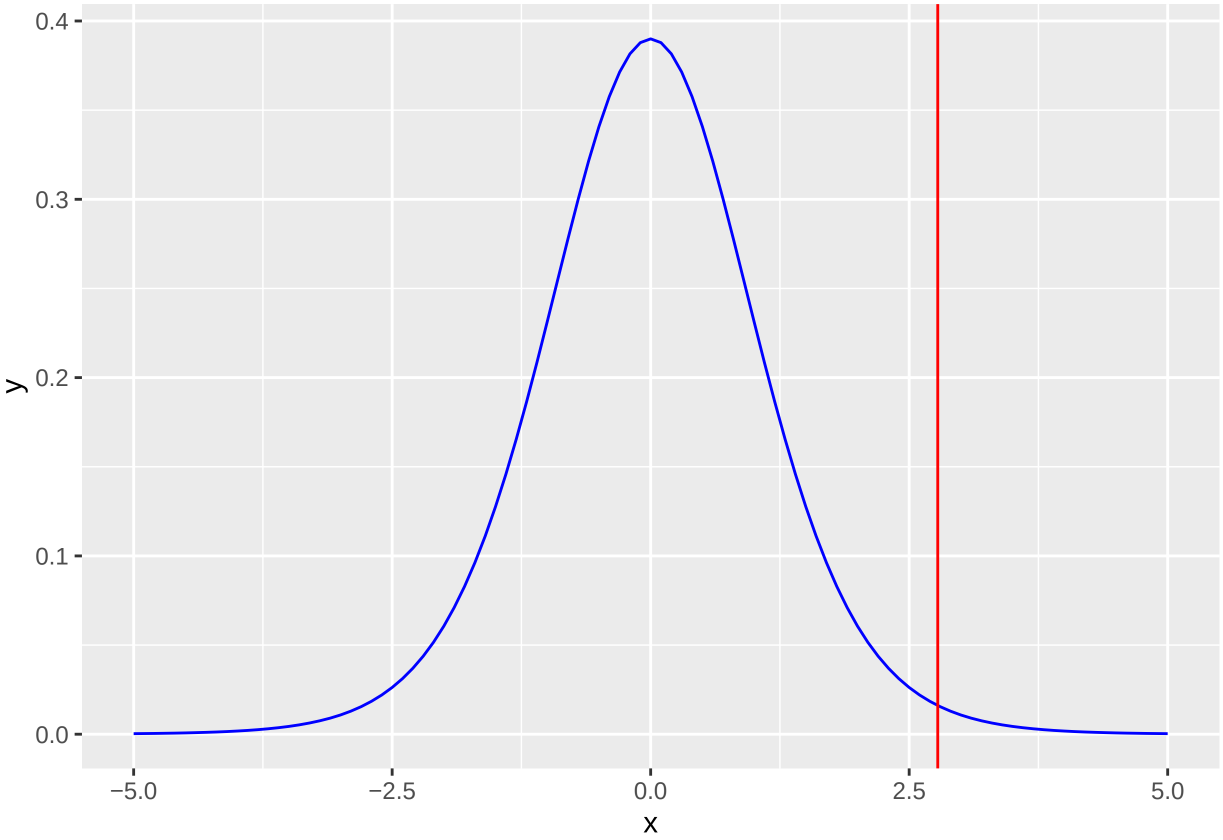
```
(SE <- sqrt((sd(s_s$`Total PCBs`)^2/nrow(s_s)) + (sd(s_n$`Total PCBs`)^2/nrow(s_n))))
```

```
## [1] 3240.948
```

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

```
## [1] 2.776451
```

```
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] 0.01801792
```

p-value is smaller than alpha significance level, thus we reject null hypothesis in favor of the alternative. There is statistically significant difference between mean PCBs in salmon from Scotland and Norway.

3. Tailor.

A high-end tailor has a detailed database of their customers measurements *data/customer_data.csv*, and a separate one with the shirt size they have been buying *data/customer_sizes.csv*.

a) Join the two datasets.

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

```
## Rows: 250 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): id, Shirt Size
##
## 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.
```

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

```
## Rows: 250 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (13): id, Age, Weight, Height, Chest, Waist Size, Hip, Thigh, Knee, Ankl...
##
## 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.
```

```
st <- customer_sizes %>% left_join(customer_data, by = 'id')
```

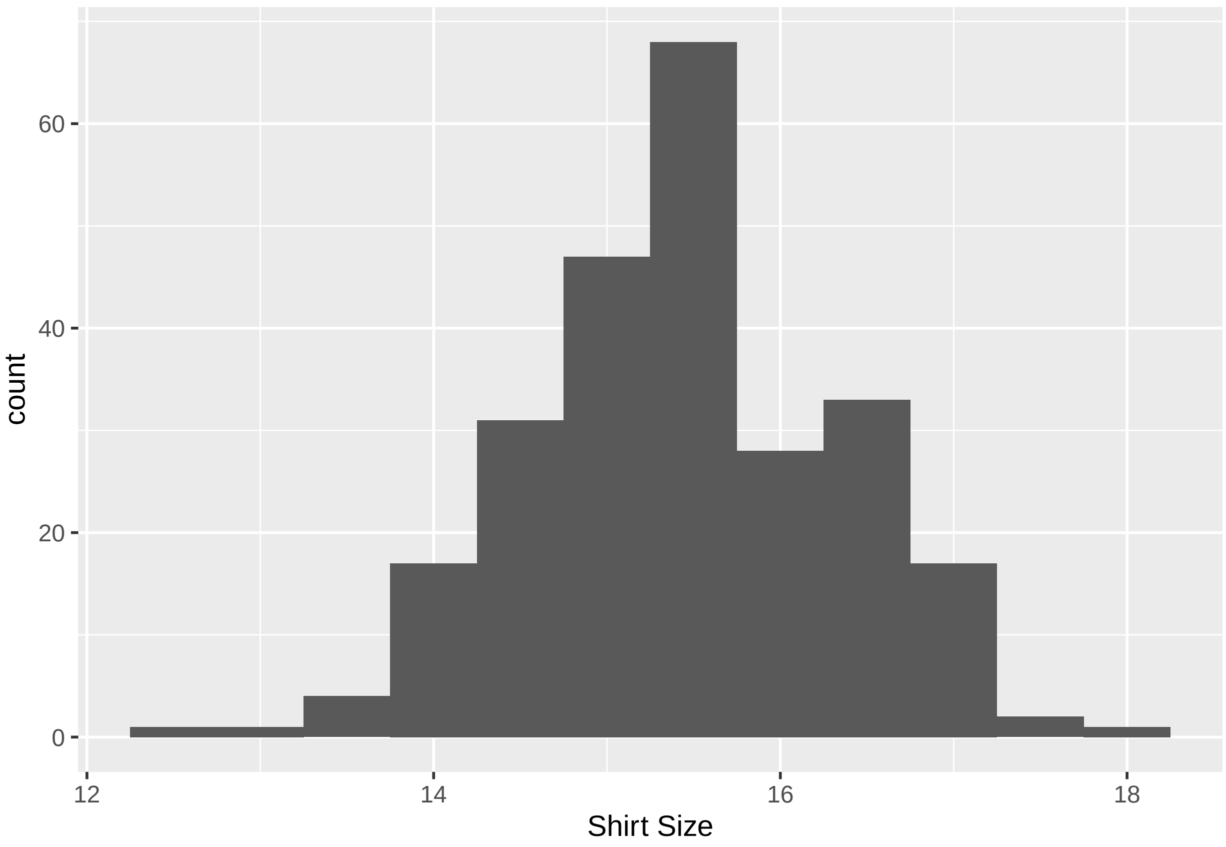
b) Which tuned multiple regression model would be suitable for predicting new customers shirt size? Create such a multiple regression model.

R2 tuned model, as it presents the best fit to the data.

```
colnames(st)
```

```
##  [1] "id"           "Shirt Size" "Age"          "Weight"      "Height"
##  [6] "Chest"        "Waist Size" "Hip"          "Thigh"       "Knee"
## [11] "Ankle"        "Bicep"      "Forearm"      "Wrist"
```

```
ggplot(st) +
  geom_histogram(aes(`Shirt Size`), binwidth=0.5)
```



No visible outliers.

R2 tuning: With all - Adjusted R-squared: 0.7108 Without Waist Size - Adjusted R-squared: 0.7121 Without Chest - Adjusted R-squared: 0.7131 No improvement after.

```
fit <- lm(`Shirt Size`~ Age +
                        Weight +
                        Height +
                        #Chest +
                        #`Waist Size` +
                        Hip +
                        Thigh +
                        Knee +
                        Ankle +
                        Bicep +
                        Forearm +
                        Wrist,
                        data = st)
summary(fit)
```

```
##
## Call:
## lm(formula = ‘Shirt Size‘ ~ Age + Weight + Height + Hip + Thigh +
##     Knee + Ankle + Bicep + Forearm + Wrist, data = st)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7168 -0.2713  0.0171  0.3148  1.4785
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.361001   1.803283   6.855 6.04e-11 ***
## Age          0.004100   0.003209   1.278  0.20251
## Weight       0.030714   0.004546   6.756 1.07e-10 ***
## Height      -0.038058   0.016553  -2.299  0.02236 *
## Hip         -0.052113   0.015756  -3.307  0.00109 **
## Thigh        0.022901   0.016177   1.416  0.15818
## Knee        -0.035401   0.027788  -1.274  0.20391
## Ankle       -0.050937   0.024642  -2.067  0.03980 *
## Bicep        0.020791   0.019163   1.085  0.27903
## Forearm      0.053940   0.023173   2.328  0.02076 *
## Wrist        0.233947   0.056797   4.119 5.25e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4847 on 239 degrees of freedom
## Multiple R-squared:  0.7246, Adjusted R-squared:  0.7131
## F-statistic:  62.9 on 10 and 239 DF,  p-value: < 2.2e-16
```

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

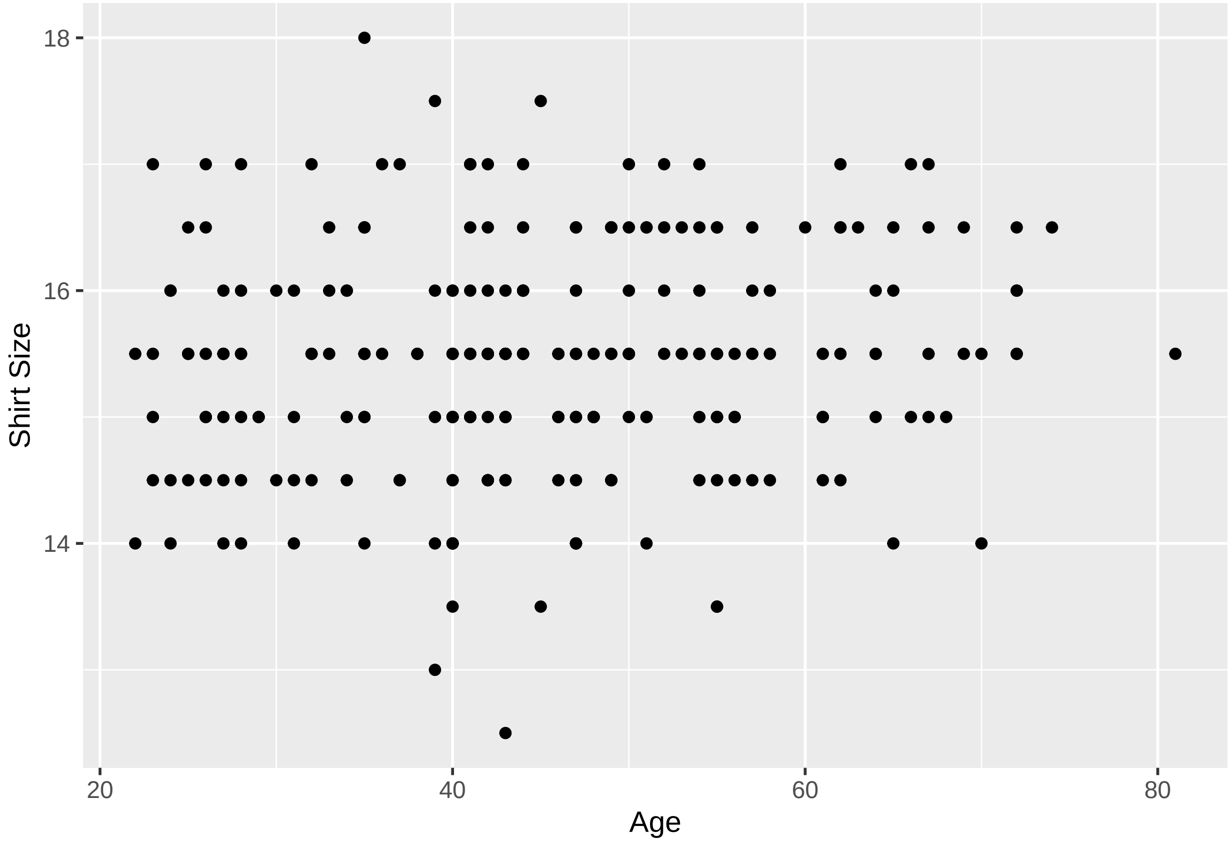
The conditions of the valid fit are:

- linearity

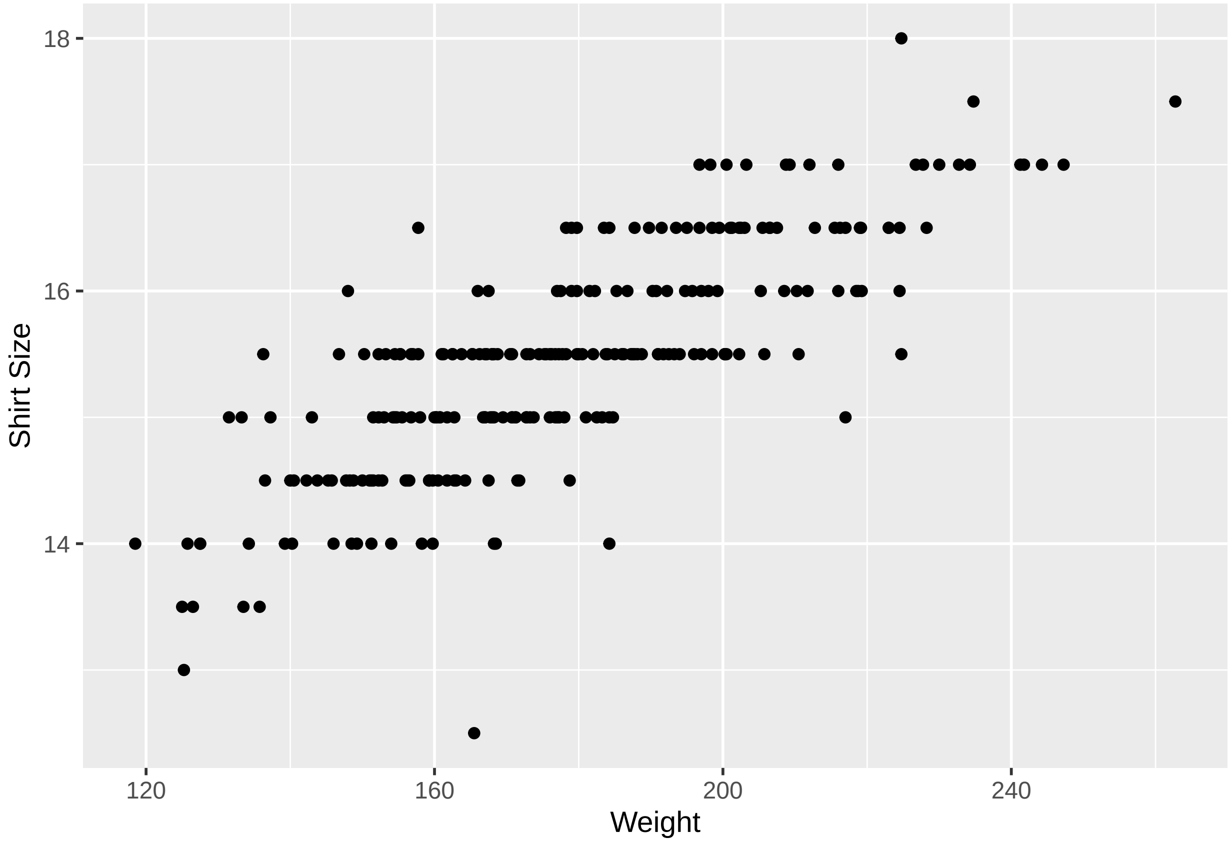
- nearly normal residuals
- constant variability
- independent observations

Linearity:

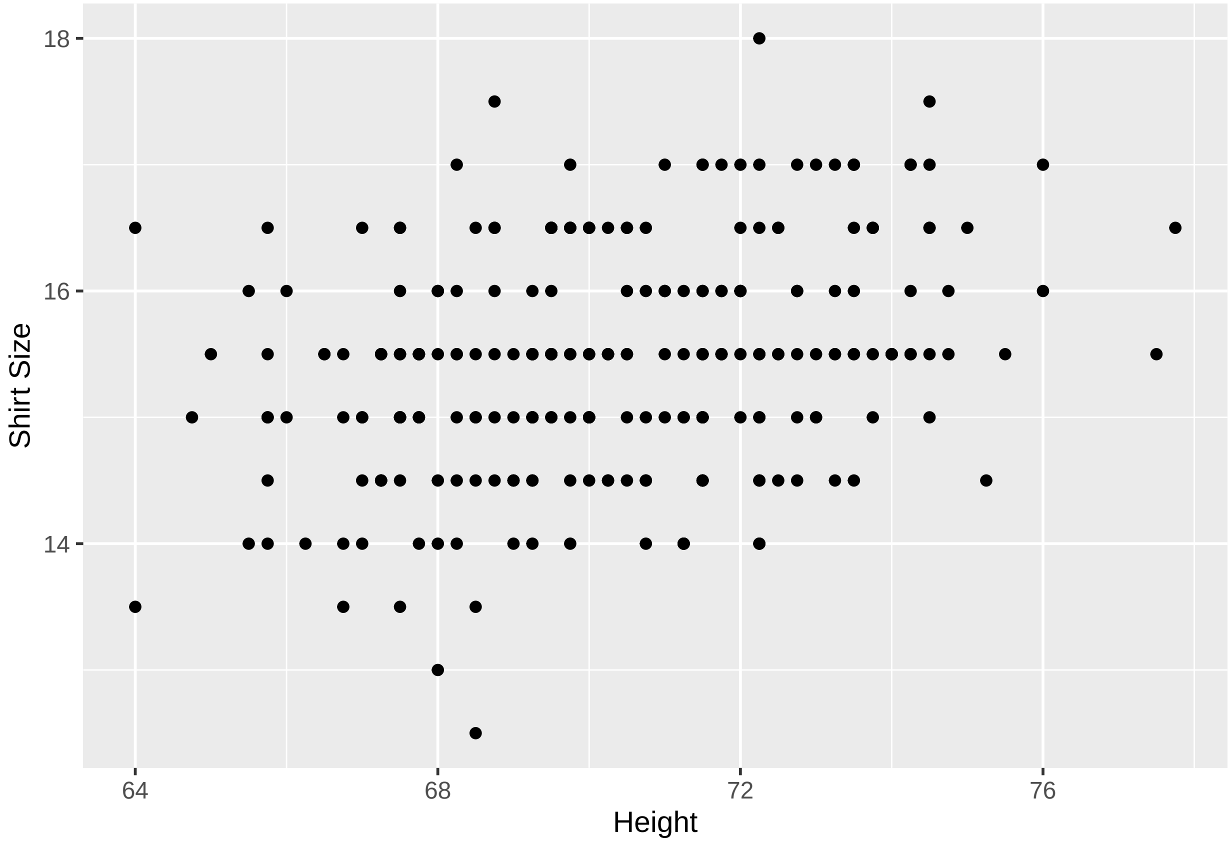
```
ggplot(st) +
  geom_point(aes(y = `Shirt Size`, x = Age))
```



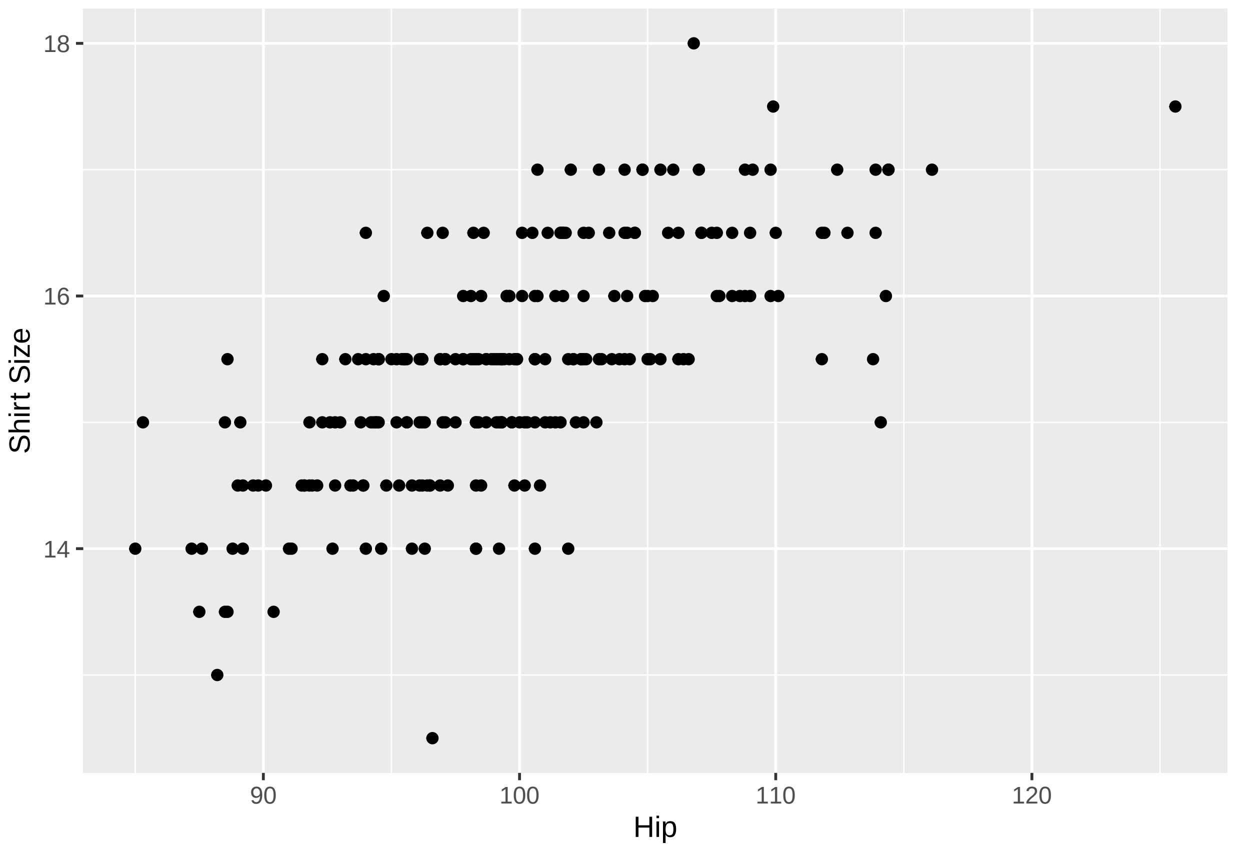
```
ggplot(st) +
  geom_point(aes(y = `Shirt Size`, x = Weight))
```



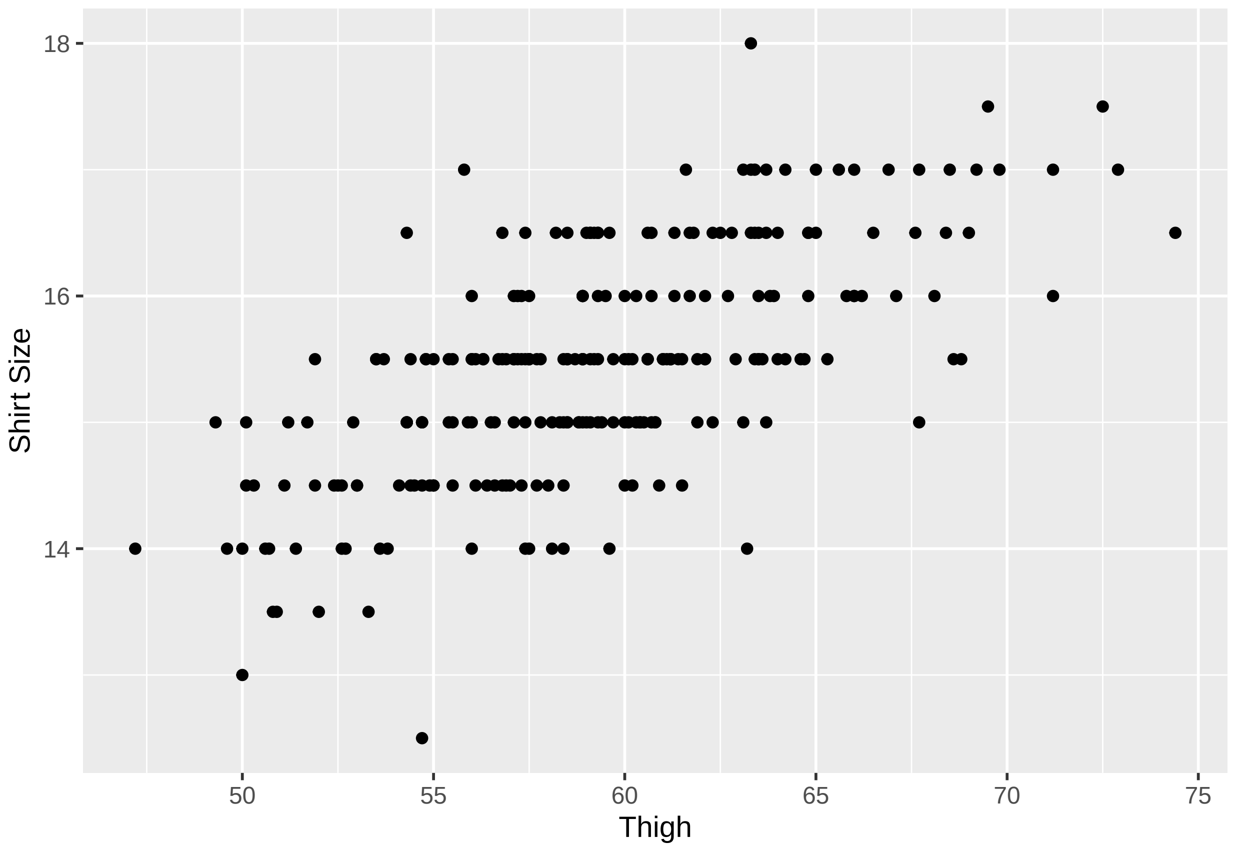
```
ggplot(st) +
  geom_point(aes(y = `Shirt Size`, x = Height))
```



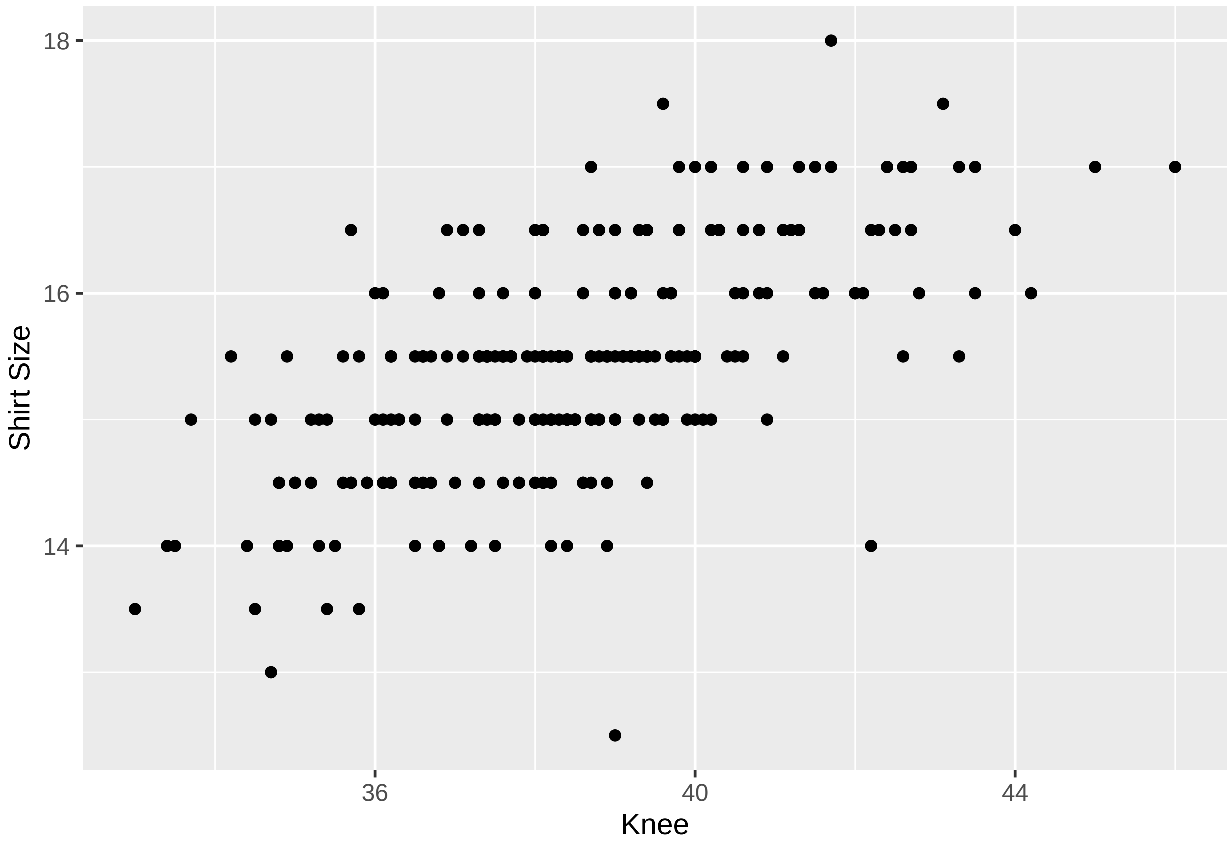
```
ggplot(st) +
  geom_point(aes(y = `Shirt Size`, x = Hip))
```



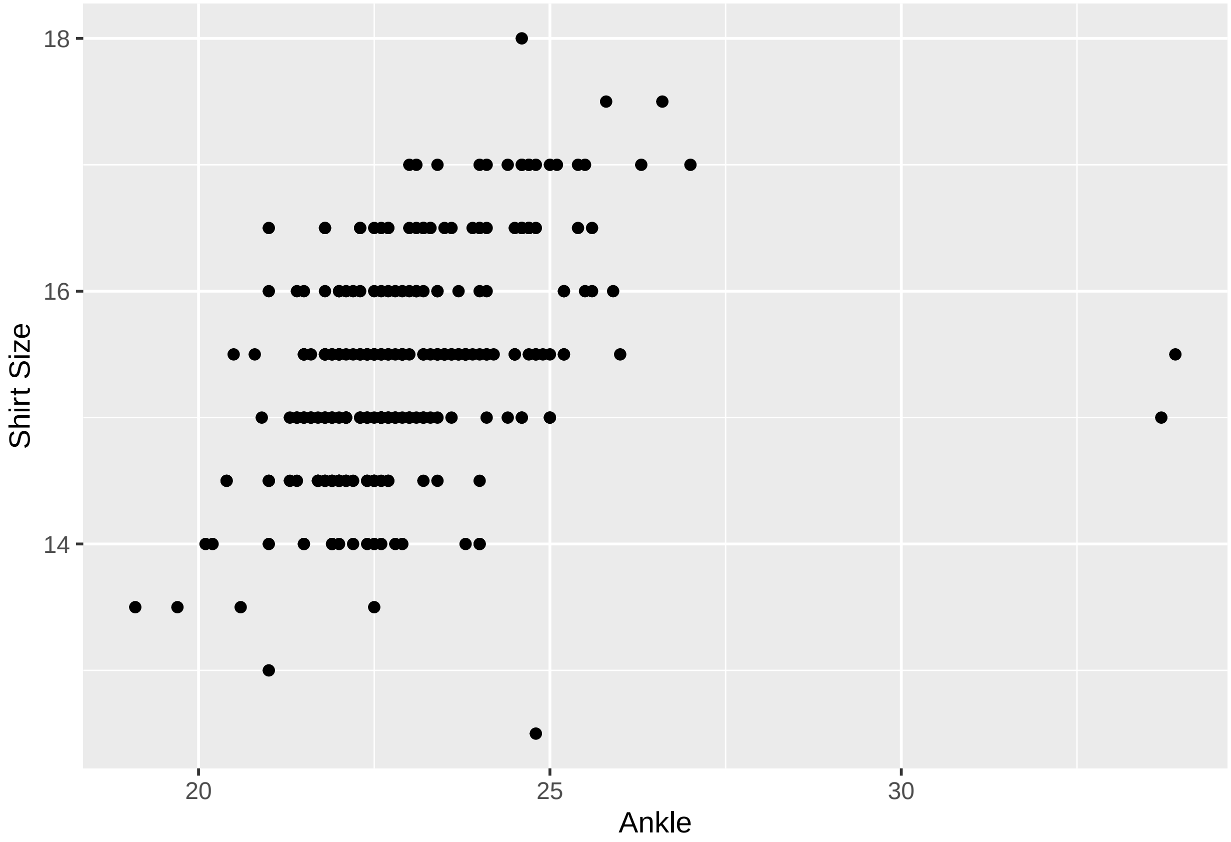
```
ggplot(st) +
  geom_point(aes(y = `Shirt Size`, x = Thigh))
```



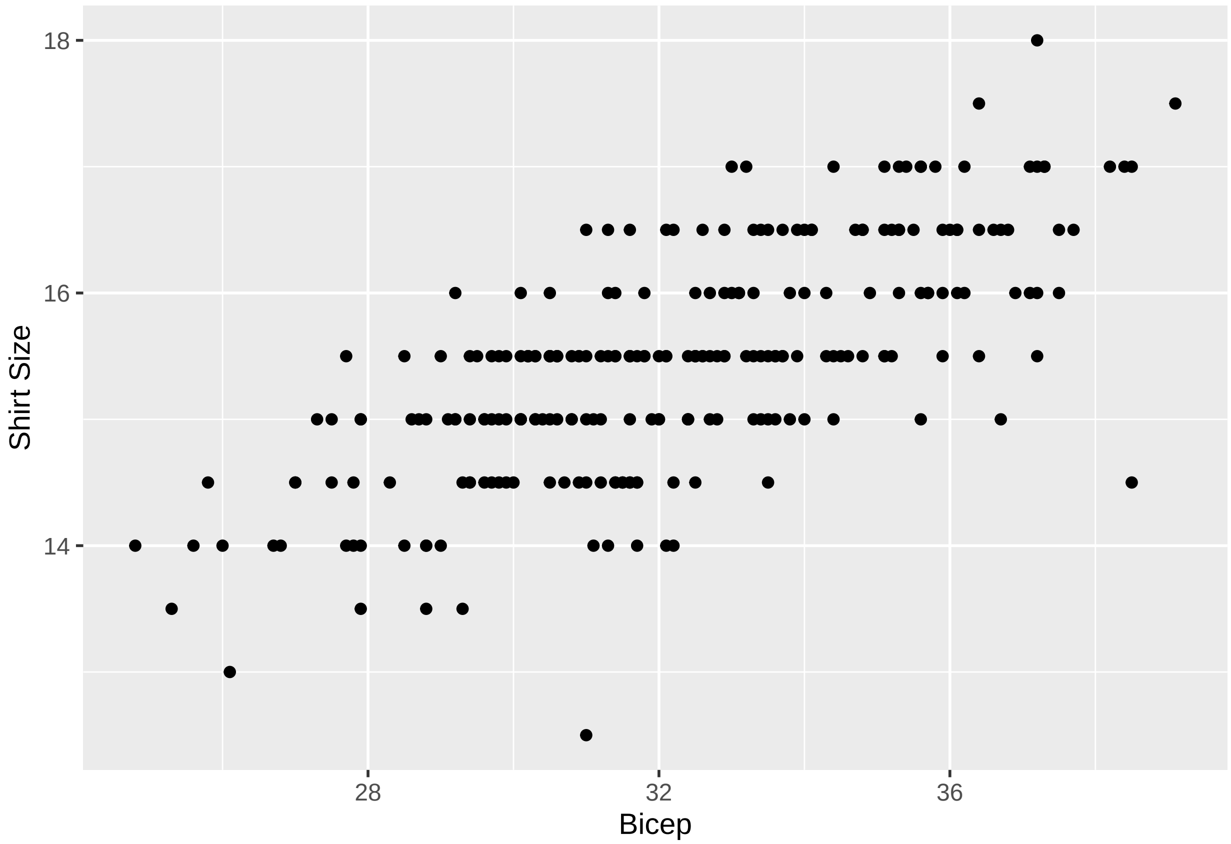
```
ggplot(st) +
  geom_point(aes(y = `Shirt Size`, x = Knee))
```



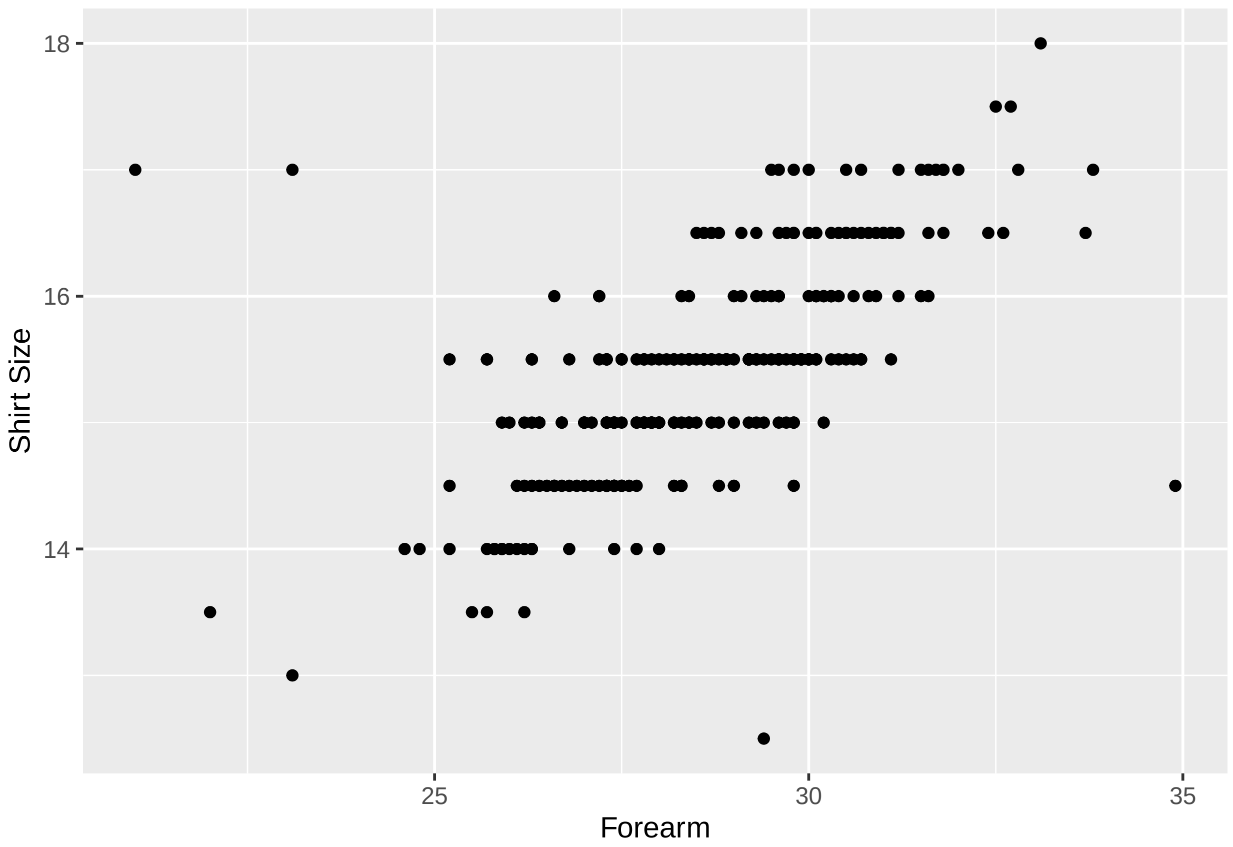
```
ggplot(st) +
  geom_point(aes(y = `Shirt Size`, x = Ankle))
```



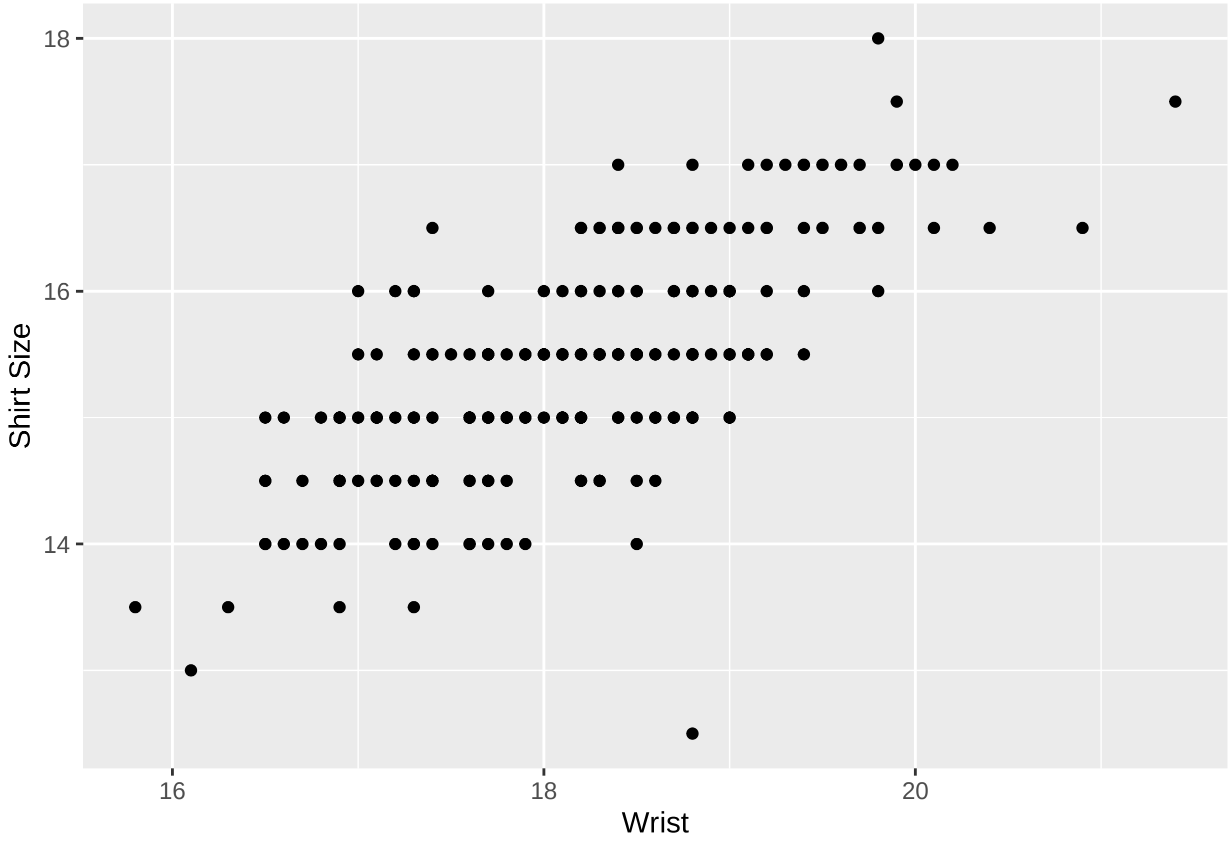
```
ggplot(st) +
  geom_point(aes(y = `Shirt Size`, x = Bicep))
```



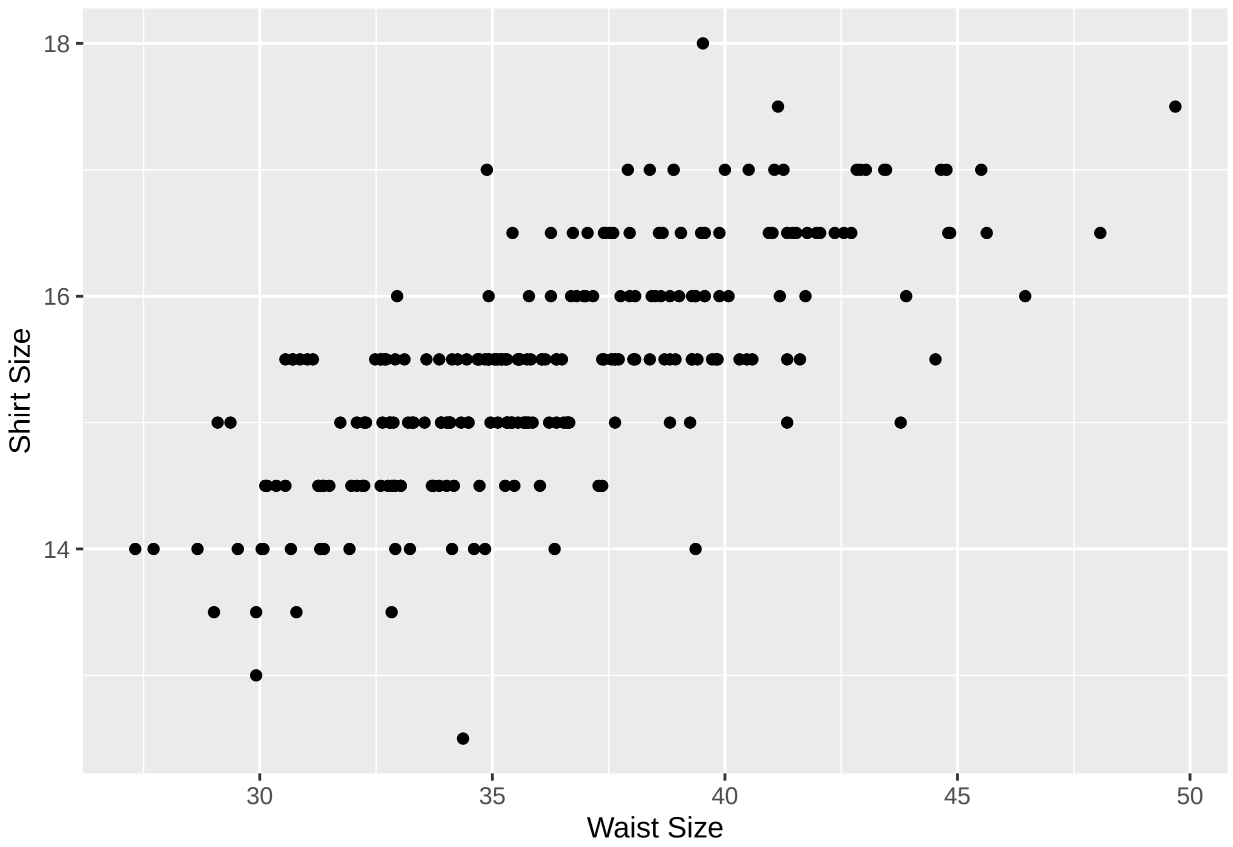
```
ggplot(st) +
  geom_point(aes(y = `Shirt Size`, x = Forearm))
```



```
ggplot(st) +
  geom_point(aes(y = `Shirt Size`, x = Wrist))
```



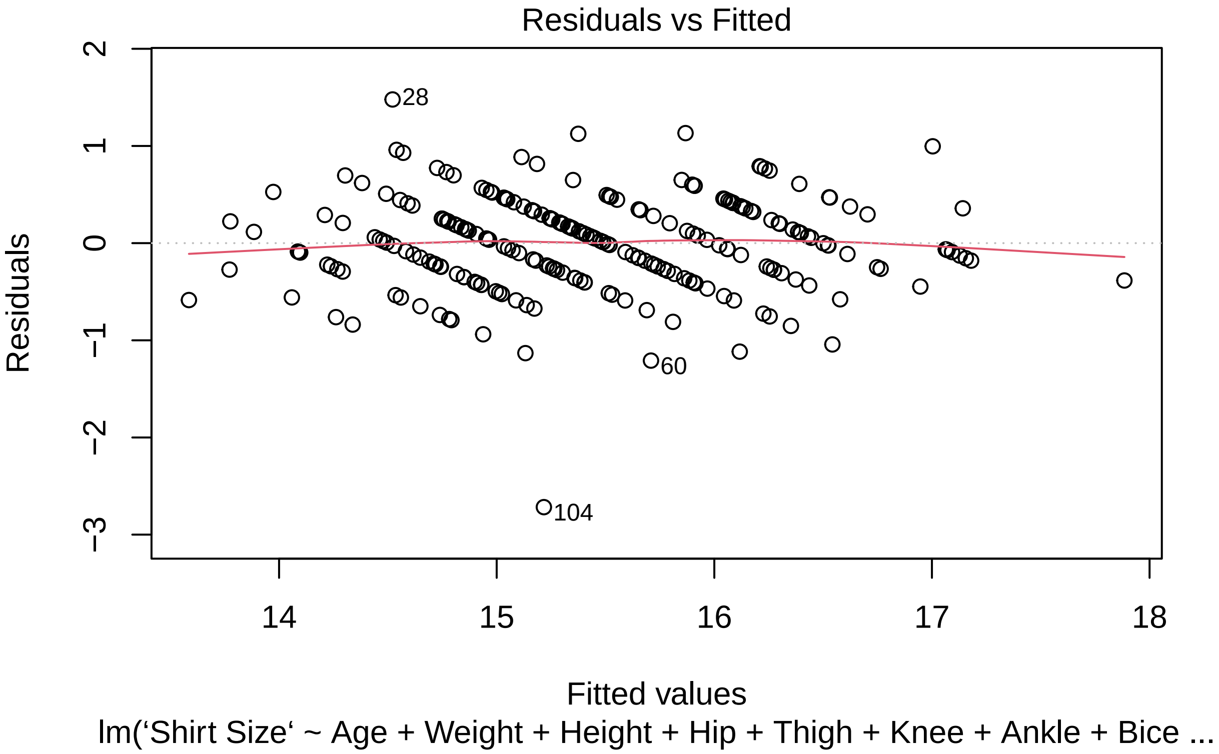
```
ggplot(st) +
  geom_point(aes(y = `Shirt Size`, x = `Waist Size`))
```

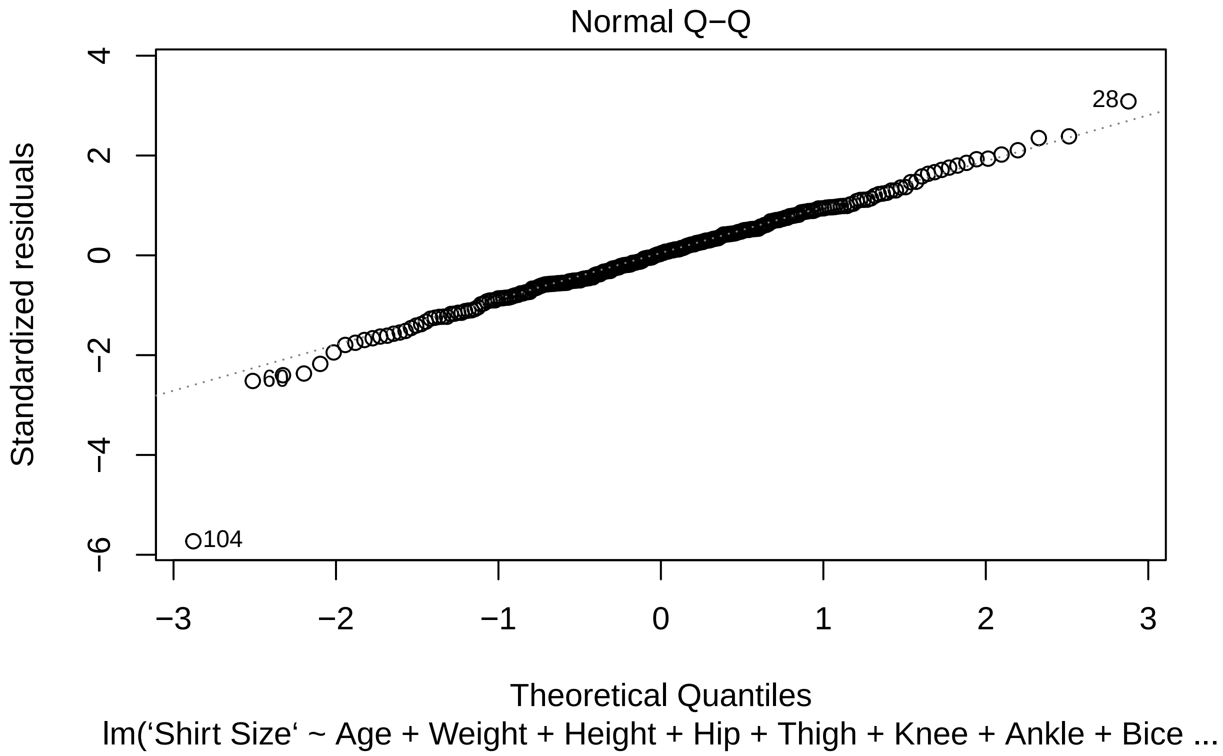


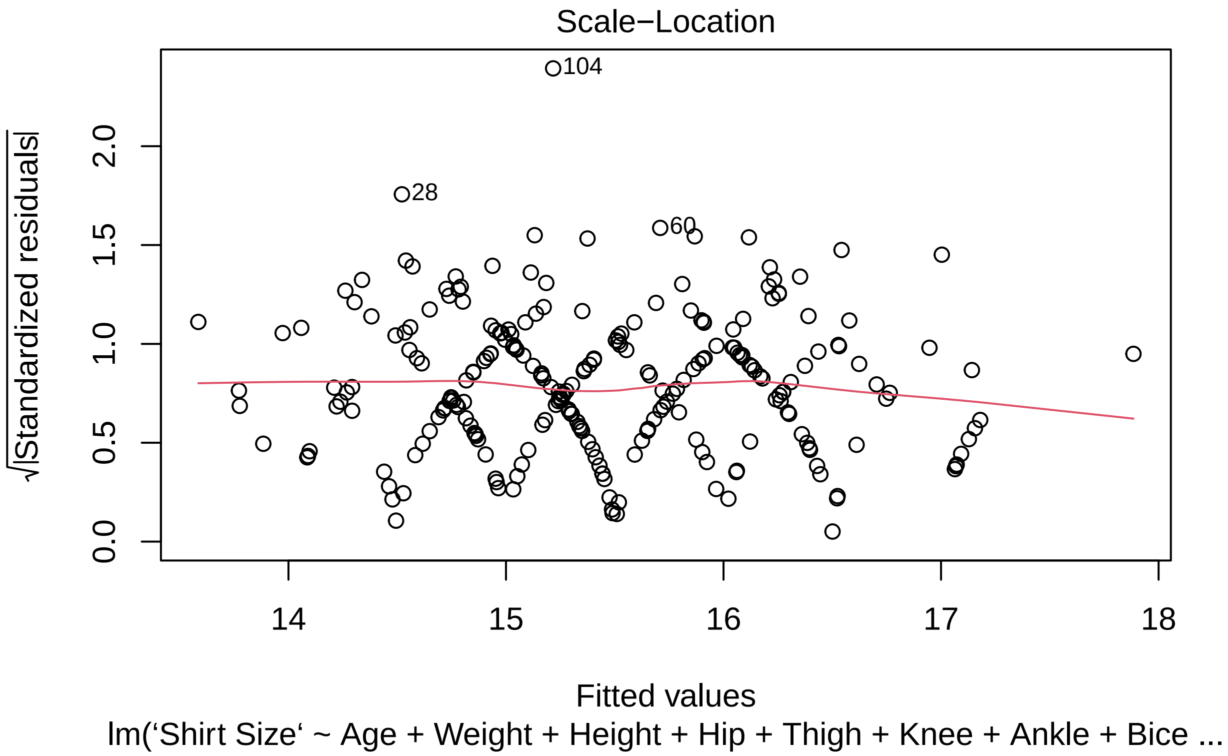
It seems there is linearity for everything except the Age.

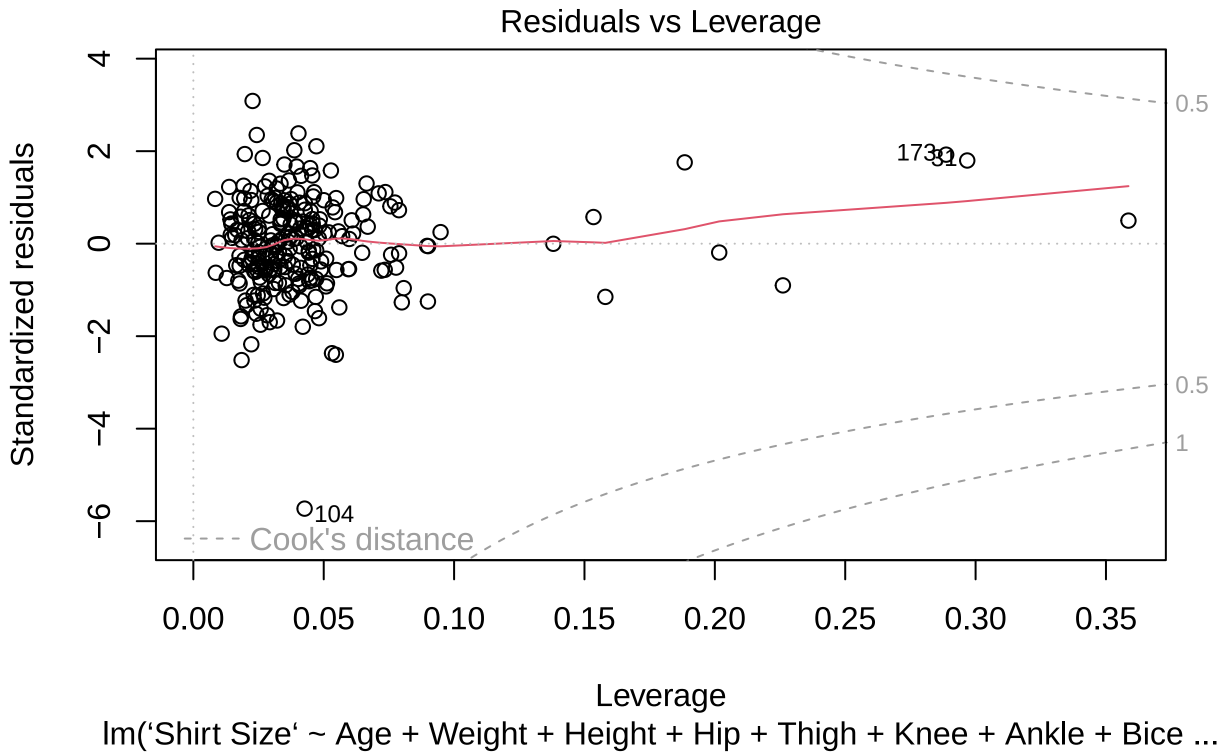
- nearly normal residuals
- constant variability

```
plot(fit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(‘Shirt Size‘ ~ Age + Weight + Height + Hip + Thigh + Knee + Ankle + Bice ...
2
1
0
−1
−2
−3
14
15
16
17
18
28
60
104


Normal Q−Q
Standardized residuals
4
2
0
−2
−4
−6
28
60
104
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(‘Shirt Size‘ ~ Age + Weight + Height + Hip + Thigh + Knee + Ankle + Bice ...

Scale−Location
104
28
60
2.0
1.5
1.0
0.5
0.0
√|Standardized residuals|
14
15
16
17
18
Fitted values
lm(‘Shirt Size‘ ~ Age + Weight + Height + Hip + Thigh + Knee + Ankle + Bice ...



There is constant variability of residuals and normal distribution. The lines happen due to the discrete nature of the ‘Shirt Size’ variables.

We assume independence of observations.

d) Predict shirt size of the new customer, Tom:

```
fit$coefficients['(Intercept)'] +
  fit$coefficients['Age'] * 43 +
  fit$coefficients['Weight'] * 136.25 +
  fit$coefficients['Height'] * 67.5 +
  fit$coefficients['Hip'] * 88.6 +
  fit$coefficients['Thigh'] * 52.0 +
  fit$coefficients['Knee'] * 34.9 +
  fit$coefficients['Ankle'] * 22.5 +
  fit$coefficients['Bicep'] * 27.7 +
  fit$coefficients['Forearm'] * 27.5 +
  fit$coefficients['Wrist'] * 18.5
```

```
## (Intercept)
##    14.73259
```

Toms size is going to be either 14.5 or 15.0.