20220103 VB-STA5 Exam Solution Guide

E. Pastucha

11/22/2021

1. North America Rodents.

Dataset *data/surveys.csv* contains information about rodents sightings in North America from 1977 to 2002. Dataset *data/species.csv* contains information about species acronyms and their Genus.

a) Join two datasets.

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

```
## New names:
## * `` -> ...1
```

```
## Rows: 30738 Columns: 10
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): species_id, sex
## dbl (8): ...1, record_id, month, day, year, plot_id, hindfoot_length, weight
```

```
##
## 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.
```

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

```
## Rows: 54 Columns: 4
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): species_id, genus, species, taxa
```

```
##
## 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.
```

```
rodents <- surveys %>% left_join(species, by = 'species_id')
```

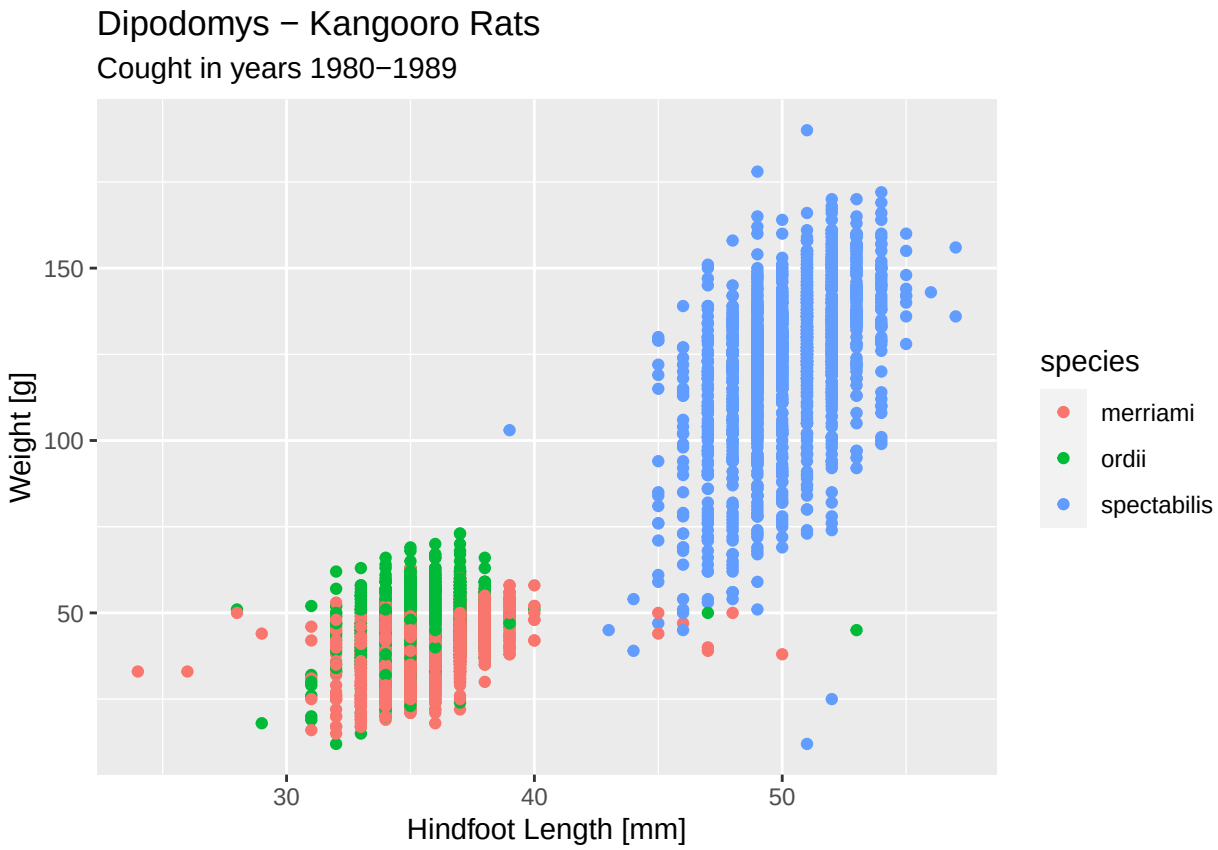
b) Present 5 heaviest rodents - their mean weight and mean hindfoot lenght in the format shown below.

```
rodents %>%
  group_by(species) %>%
  summarise(`Mean Weight [g]` = mean(weight), `Mean Hindfoot Length [mm]` = mean(hindfoot_length)) %>%
  arrange(desc(`Mean Weight [g]`)) %>%
  head(5) %>%
  knitr::kable()
```

species	Mean Weight [g]	Mean Hindfoot Length [mm]
albigula	158.72371	32.25048
spectabilis	120.21668	49.99260
hispidus	64.84906	28.05031
fulviventer	59.12500	26.70000
ochrognathus	55.37500	25.60000

c) Recreate the plot.

```
rodents %>%
  filter(year >= 1980 & year < 1990) %>%
  filter(genus == 'Dipodomys') %>%
  ggplot() +
  geom_point(aes(y = weight, x = hindfoot_length, color = species)) +
  labs(title = 'Dipodomys - Kangooro Rats',
       subtitle = 'Cought in years 1980-1989',
       x = 'Hindfoot Length [mm]',
       y = 'Weight [g]')
```



d) Describe the plot.

- The plot presents relationship between Kangooro Rats hindfoot length and their weight.
- There are 3 species within this genom. Merriami, ordii and spectabilis.
- Spectabilis species is heavier and has bigger feet than the other species.
- Ordii seems slightly heavier than merriami.
- There are multiple outliers - maybe due to mislabeling.
- There seems to be linear relationship in between 2 variables, but not very strong.

e) Kangooro Rats (genom *Dipodomys*) are small rodents moving similarily to kangooros using jumping steps. Is the mean male hindfoot length different between *ordii* and *merriami* species. Comment on the results.

Difference of means t-test.

$H_0 : \mu_{ordii} - \mu_{merriami} = 0$

$H_A : \mu_{ordii} - \mu_{merriami} = 0 \neq 0$

H0: There is no difference between mean ordii and mean merriami mean hindfoot length.

HA: There is difference between mean ordii and mean merriami mean hindfoot length.

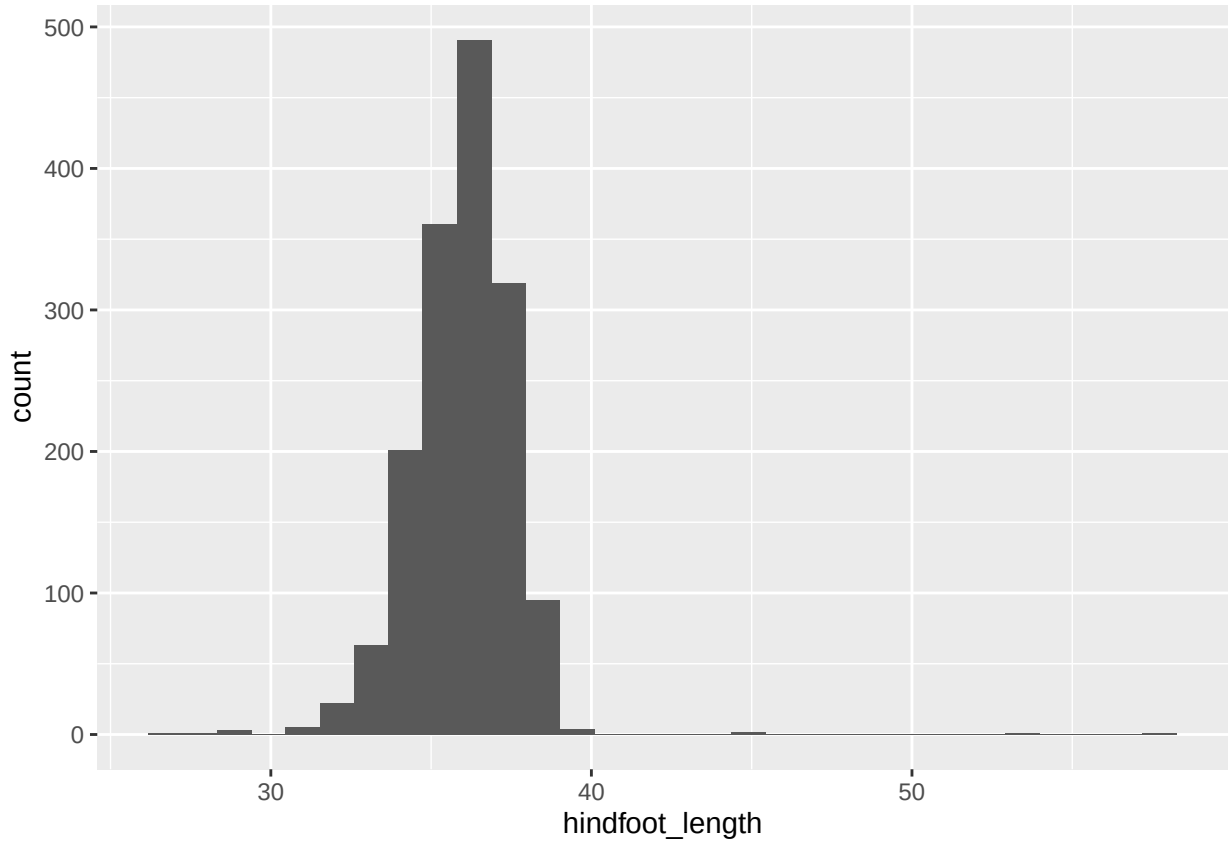
alpha significance level - 0.05

Conditions check:

Normality:

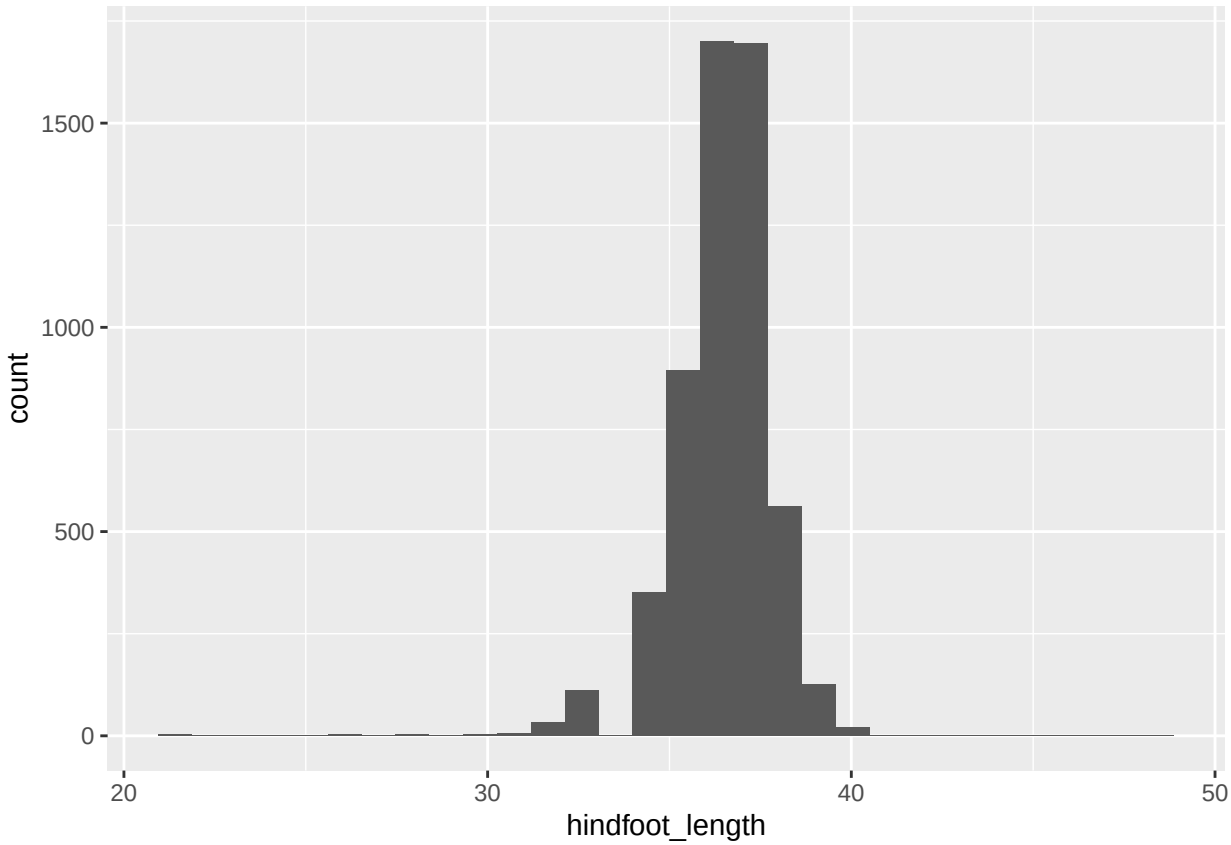
```
rodents %>%
  filter(species == 'ordii') %>%
  filter(sex == 'M') %>%
  ggplot() +
  geom_histogram(aes(x = hindfoot_length))
```

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




```
rodents %>%
  filter(species == 'merriami') %>%
  filter(sex == 'M') %>%
  ggplot() +
  geom_histogram(aes(x = hindfoot_length))
```

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



There are multiple outliers, but it seems that most variables follow normal distribution.

We assume that observations are independent.

- short version

```
rodents %>%
  filter(species == 'ordii' | species == 'merriami') %>%
  filter(sex == 'M') %>%
  t.test(hindfoot_length~species, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  hindfoot_length by species
## t = 11.463, df = 2333.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4302448 0.6078253
## sample estimates:
## mean in group merriami    mean in group ordii
##               36.19674               35.67771
```

p-value is smaller than alpha significance level, thus we reject null hypothesis and accept the alternative. The mean male hindfoot length in between ordii and merriami species is not the same. Howeve rhe difference is 0.5 mm, which is statistically significant but practically insignificant.

- long version

```
ordii <- rodents %>%
  filter(species == 'ordii') %>%
  filter(sex == 'M')
merriami <- rodents %>%
  filter(species == 'merriami') %>%
  filter(sex == 'M')

(point_estimate <- mean(ordii$hindfoot_length) - mean(merriami$hindfoot_length))
```

```
## [1] -0.5190351
```

```
nrow(ordii)
```

```
## [1] 1570
```

```
nrow(merriami)
```

```
## [1] 5525
```

```
dof <- 1569
```

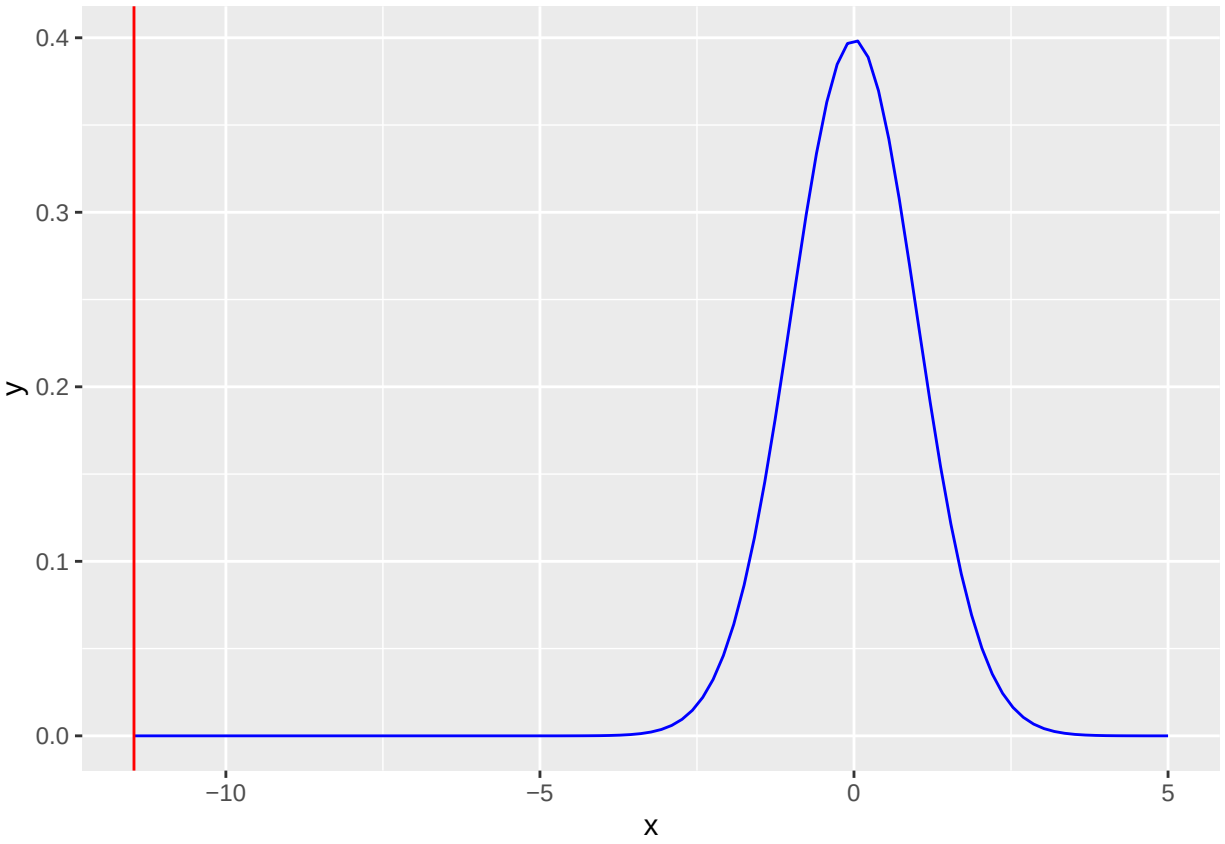
```
(SE <- sqrt((sd(ordii$hindfoot_length)^2/nrow(ordii)) + (sd(merriami$hindfoot_length)^2/nrow(merriami))
```

```
## [1] 0.04527848
```

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

```
## [1] -11.46317
```

```
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 <- dt(t_score, df = dof))
```

```
## [1] 1.518346e-28
```

p-value is smaller than alpha significance level, thus we reject null hypothesis and accept the alternative. The mean male hindfoot length in between ordii and merriami species is not the same. Howeve rhe difference is 0.5 mm, which is statistically significant but practically insignificant.

2. Medical students smoking habits.

A study was conducted on various Medical Universities within Germany and Hungary. The students were asked about their smoking habits. 2883 students took part, *44%* of them were German, *36%* were Hungarian and *20%* other nationalities. The table below lists number of students per nationality declaring daily smoking habit.

Nationality	n
German	91
Hungarian	78
Multinational	51

a) Are the proportions of nationalities within smoking students a true representation of proportions of whole student body? Conduct a suitable test to check this hypothesis.

Chi square test for goodness of fit.

H0: Distribution of nationalities within smoking students is the same as distribution of whole student body.

HA: Distribution of nationalities within smoking students is not the same as distribution of whole student body.

alpha significance level - 0.05

Conditions check:

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

```
all_smoking <- 91+78+51
(expected_G <- 0.44 * all_smoking)
```

```
## [1] 96.8
```

```
(expected_H <- 0.36 * all_smoking)
```

```
## [1] 79.2
```

```
(expected_M <- 0.20 * all_smoking)
```

```
## [1] 44
```

All expected values are above 5.

- short version

```
chisq.test(c(91,78,51), p=c(0.44,0.36,0.20))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(91, 78, 51)
## X-squared = 1.4793, df = 2, p-value = 0.4773
```

We accept null hypothesis. The distribution of smoking students nationalities is not significantly different from all students nationalities distribution.

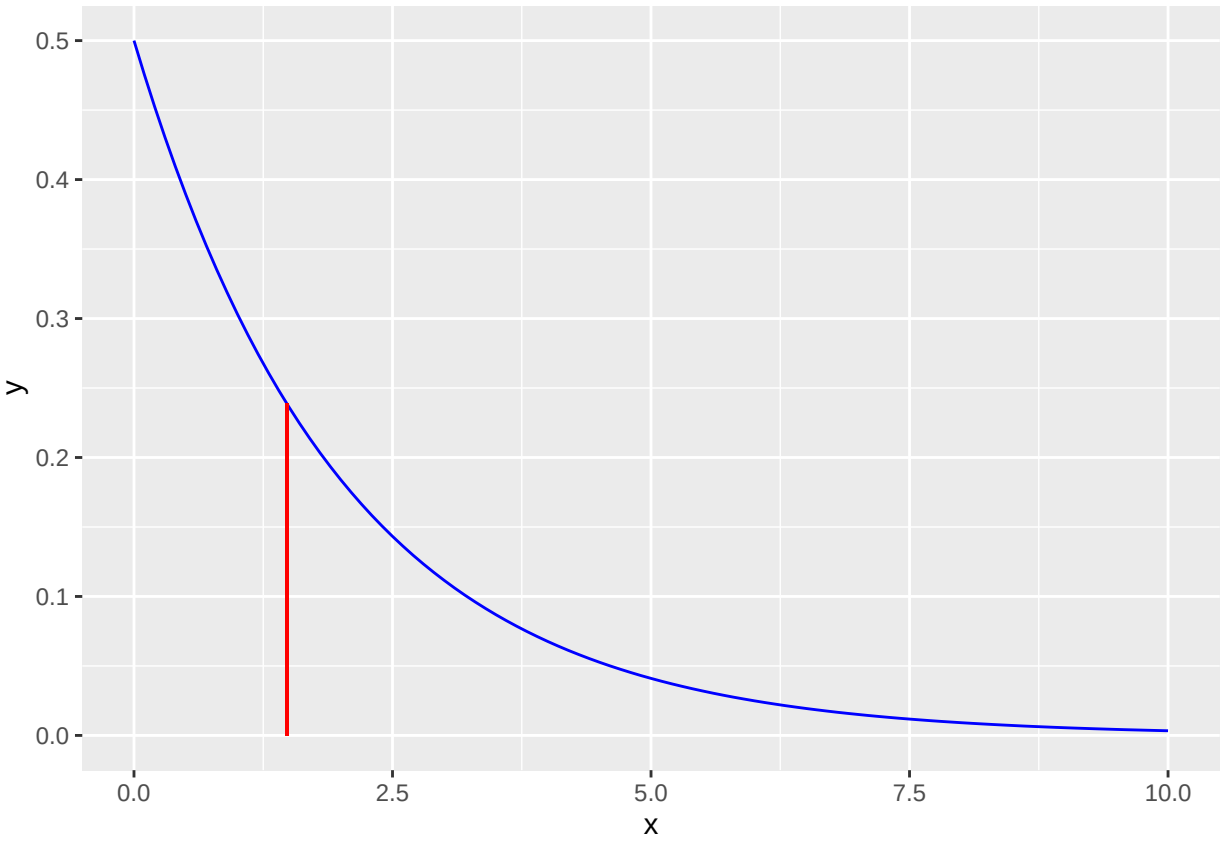
- long version

```
(chi2_stat <- ((91 - expected_G)^2/expected_G) +
              ((78 - expected_H)^2/expected_H) +
              ((51 - expected_M)^2/expected_M))
```

```
## [1] 1.479339
```

```r
dof <- 2
```

```r
ggplot(data.frame(x = seq(0, 10, length=100)), aes(x = x)) +
  stat_function(fun = dchisq, args = list(df = dof), color = 'blue') +
  geom_segment(aes(x = chi2_stat, y = 0, xend = chi2_stat, yend = dchisq(chi2_stat, df = dof)),  color =
```




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

```
## [1] 0.4772717
```

We accept null hypothesis. The distribution of smoking students nationalities is not significantly different from all students nationalities distribution.

3. University salaries.

Dataset *data/salaries.csv* contains data about yearly salaries of random 52 academic workers at one of the U.S. Universities. Your friend has been working there for the past *10 years*. He achieved his doctorate *12 years ago*, and is an *associate professor* in the Geology Department. He wants to know, if his salary is appropriate, higher than expected, or lower than expected.

a) Create a model to predict salaries at this University. Tune it, so that it is most statisticaly significant.

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

```
## Rows: 52 Columns: 6

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): sx, rk, dg
## dbl (3): yr, yd, sl

##
## 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
colnames(salaries)
```

```
## [1] "sx" "rk" "yr" "dg" "yd" "sl"
```

```r
fit <- lm(sl~ sx +
            rk +
            dg +
            yr +
            yd,
          data = salaries)
summary(fit)
```

```
##
## Call:
## lm(formula = sl ~ sx + rk + dg + yr + yd, data = salaries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4045.2 -1094.7  -361.5   813.2  9193.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16912.42     816.44  20.715  < 2e-16 ***
## sxmale      -1166.37     925.57  -1.260    0.214
## rkassociate  5292.36    1145.40   4.621 3.22e-05 ***
## rkfull      11118.76    1351.77   8.225 1.62e-10 ***
## dgmasters    1388.61    1018.75   1.363    0.180
## yr            476.31      94.91   5.018 8.65e-06 ***
## yd           -124.57      77.49  -1.608    0.115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2398 on 45 degrees of freedom
## Multiple R-squared:  0.855,  Adjusted R-squared:  0.8357
## F-statistic: 44.24 on 6 and 45 DF,  p-value: < 2.2e-16
```

```r
fit <- lm(sl~ #sx +
            rk +
            dg +
            yr +
            yd,
          data = salaries)
summary(fit)
```

```
##
## Call:
## lm(formula = sl ~ rk + dg + yr + yd, data = salaries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4226.9  -972.1  -293.1   612.5  9840.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16311.34     666.84  24.461  < 2e-16 ***
## rkassociate  4713.92    1056.09   4.464 5.18e-05 ***
## rkfull      10509.62    1270.43   8.272 1.18e-10 ***
## dgmasters    1062.12     991.53   1.071    0.290
## yr            416.56      82.75   5.034 7.84e-06 ***
## yd            -81.22      69.87  -1.162    0.251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2414 on 46 degrees of freedom
## Multiple R-squared:  0.8499, Adjusted R-squared:  0.8336
## F-statistic:  52.1 on 5 and 46 DF,  p-value: < 2.2e-16
```

```r
fit <- lm(sl~ #sx +
            rk +
            #dg +
            yr +
            yd,
          data = salaries)
summary(fit)
```

```
##
## Call:
## lm(formula = sl ~ rk + yr + yd, data = salaries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3413.7 -1218.5  -182.7   742.0  9483.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16317.46     667.86  24.433  < 2e-16 ***
## rkassociate  4619.12    1054.02   4.382 6.55e-05 ***
## rkfull       9864.30    1120.27   8.805 1.65e-11 ***
```

```
## yr            400.46      81.50   4.914 1.13e-05 ***
## yd            -34.32      54.54  -0.629    0.532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2417 on 47 degrees of freedom
## Multiple R-squared:  0.8462, Adjusted R-squared:  0.8331
## F-statistic: 64.64 on 4 and 47 DF,  p-value: < 2.2e-16
```

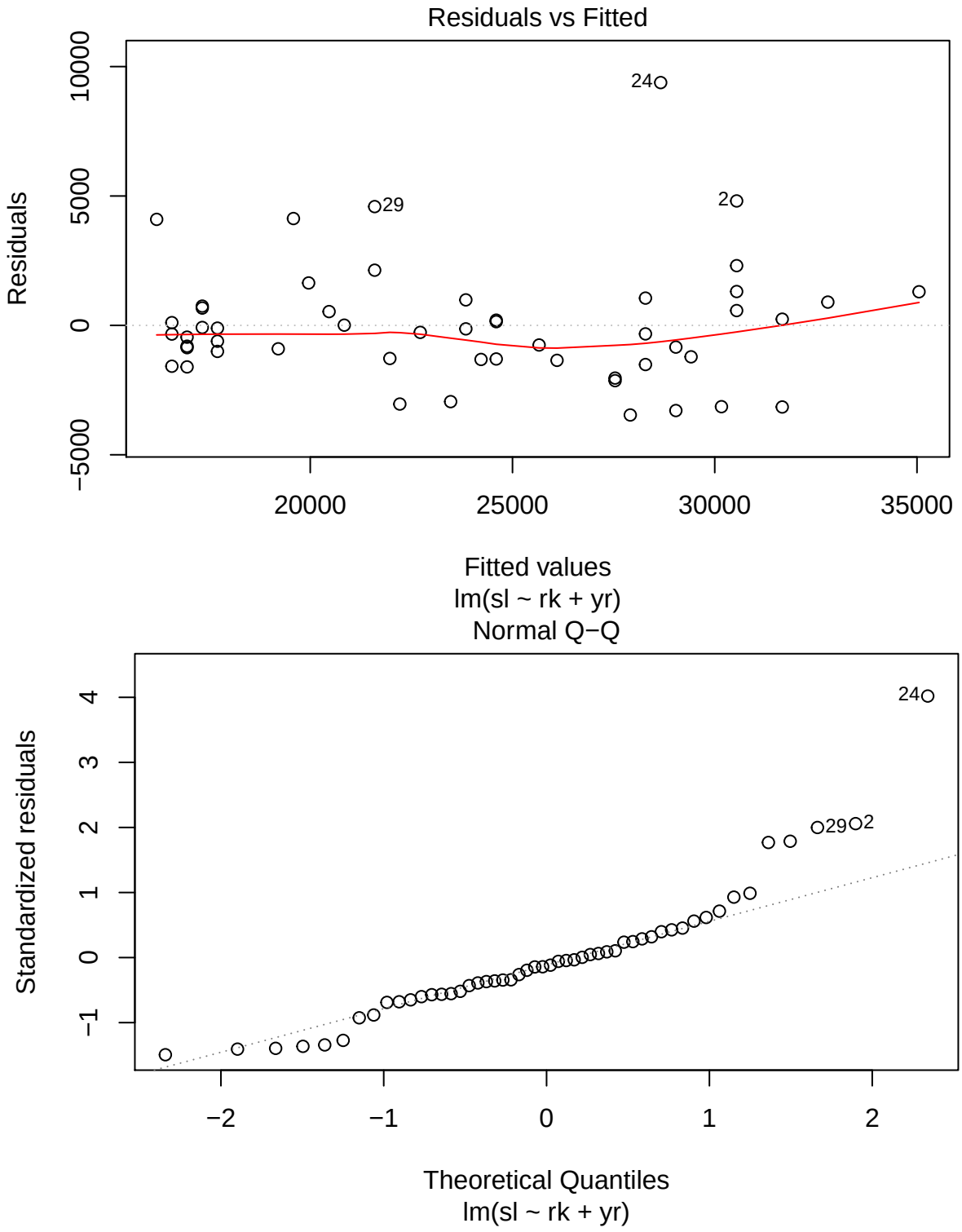
```
fit <- lm(sl~ #sx +
            rk +
            #dg +
            yr #+
            #yd
          ,data = salaries)
summary(fit)
```

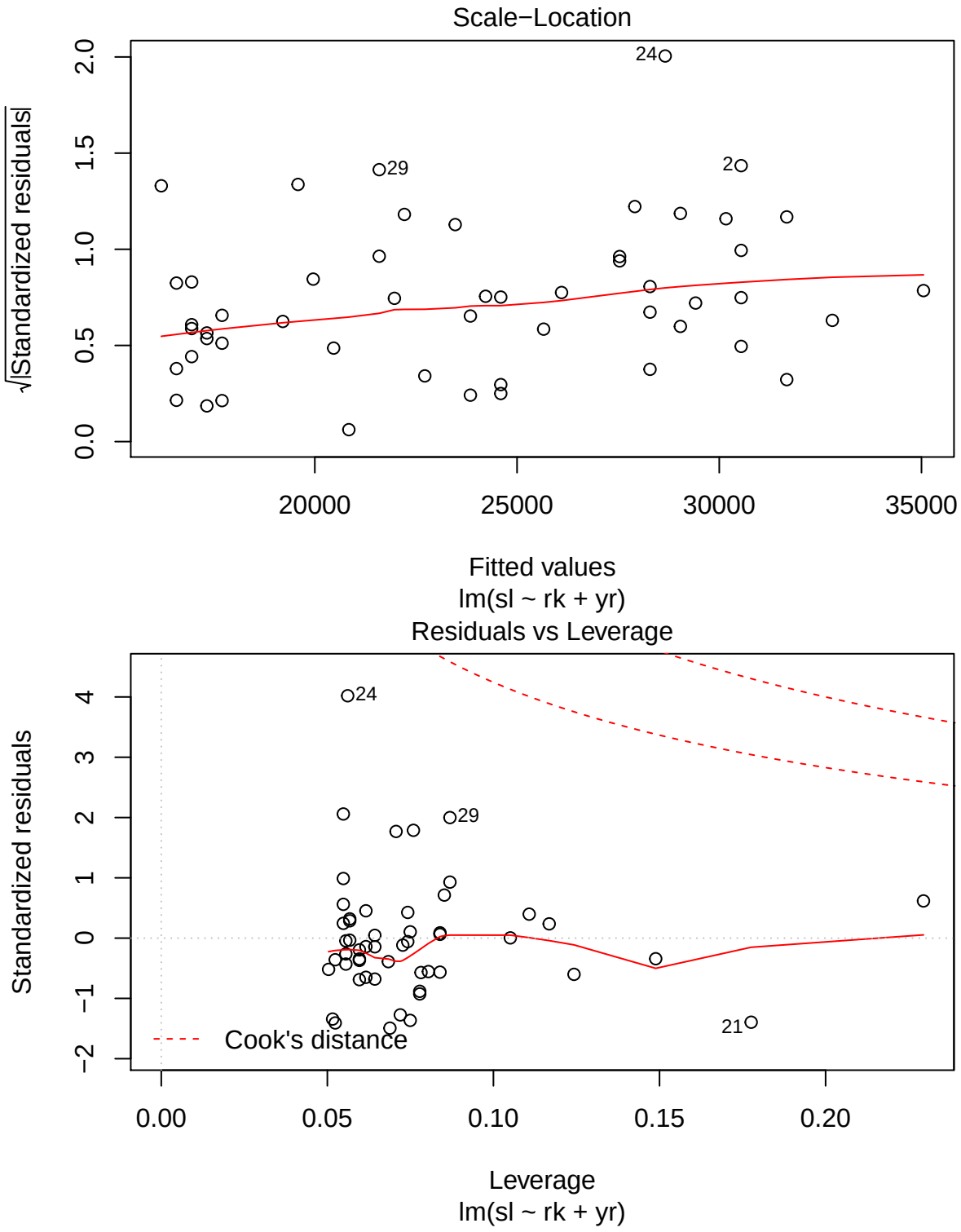
```
##
## Call:
## lm(formula = sl ~ rk + yr, data = salaries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3462.0 -1302.8  -299.2   783.5  9381.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16203.27     638.68  25.370  < 2e-16 ***
## rkassociate  4262.28     882.89   4.828 1.45e-05 ***
## rkfull       9454.52     905.83  10.437 6.12e-14 ***
## yr            375.70      70.92   5.298 2.90e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2402 on 48 degrees of freedom
## Multiple R-squared:  0.8449, Adjusted R-squared:  0.8352
## F-statistic: 87.15 on 3 and 48 DF,  p-value: < 2.2e-16
```

b) Are the conditions for a valid linear fit fulfilled in this case?

- constant residuals and normal distribution of residuals

```
plot(fit)
```

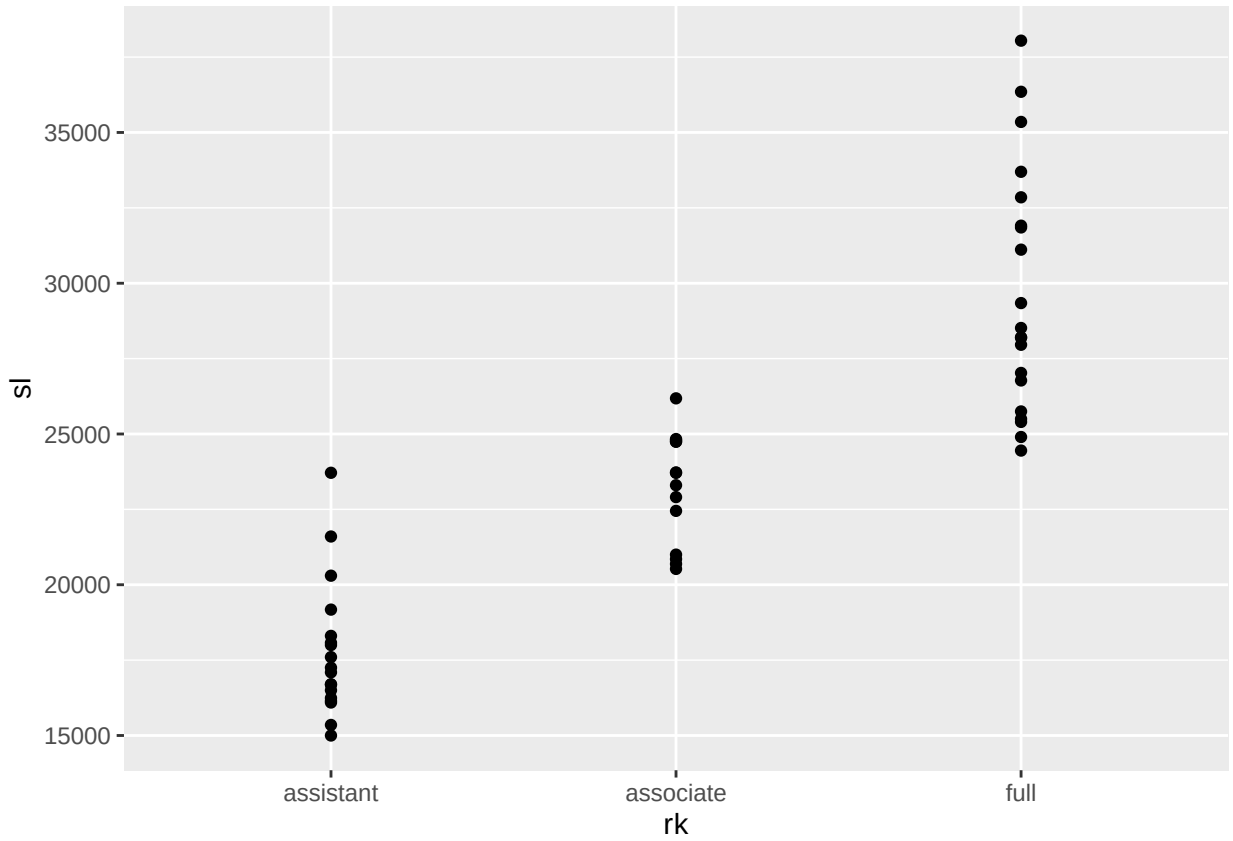

Residuals vs Fitted
Residuals
10000
5000
0
−5000
24
29
2
20000
25000
30000
35000
Fitted values
lm(sl ~ rk + yr)
Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
24
29
2
−2
−1
0
1
2
Theoretical Quantiles
lm(sl ~ rk + yr)



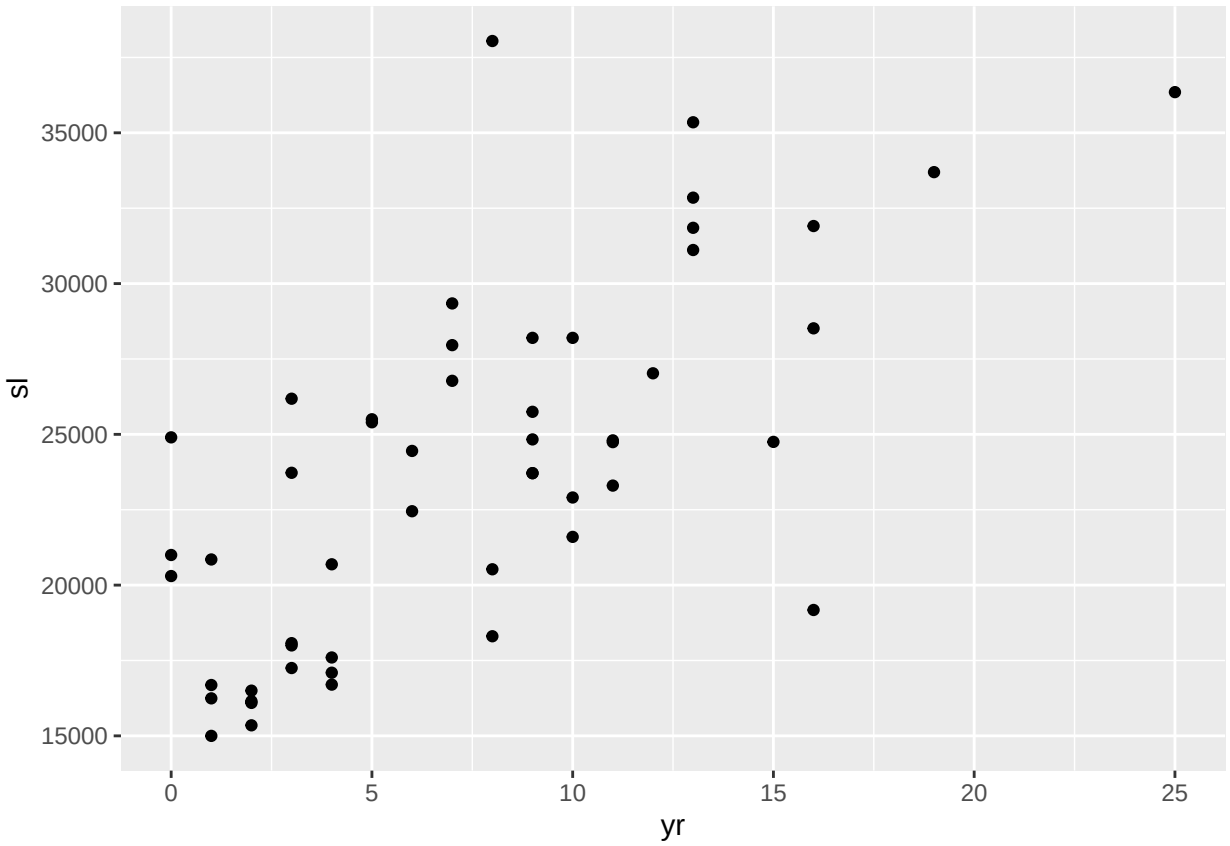
As seen at the first two plots, there are some outliers from normal distribution of the residuals, however majority of points follow theoretical line. The residuals also seem to be evanly distributed.

- we assume independence
- linearity

```
ggplot(salaries) +
  geom_point(aes(x = rk, y = sl))
```




```
ggplot(salaries) +
  geom_point(aes(x = yr, y = sl))
```



There seems to be some linear corelation visible on the plots.

c) Predict your friends' salary. Help him to make a decission - contact HR or stay quiet.

```
(predicted_salary <- 16203.27 + 4262.28 + 375.70*10)
```

```
## [1] 24222.55
```

This friend earns more than predicted in the model. I would not contact HR.