Introductory Statistics with R

SCSRU Workshop

Statistical Consulting and Survey Research Unit
University of Waterloo

2024-11-13

Dataset Overview

Throughout this workshop, we will use the `pokemon` dataset.

Dataset Overview

```r
str(df)
```

```
## 'data.frame':    801 obs. of  11 variables:
##  $ name        : chr  "Bulbasaur" "Ivysaur" "Venusaur" "Charmander" ...
##  $ weight_kg   : num  6.9 13 100 8.5 19 90.5 9 22.5 85.5 2.9 ...
##  $ height_m    : num  0.7 1 2 0.6 1.1 1.7 0.5 1 1.6 0.3 ...
##  $ hp          : int  45 60 80 39 58 78 44 59 79 45 ...
##  $ attack      : int  49 62 100 52 64 104 48 63 103 30 ...
##  $ defense     : int  49 63 123 43 58 78 65 80 120 35 ...
##  $ sp_attack   : int  65 80 122 60 80 159 50 65 135 20 ...
##  $ sp_defense  : int  65 80 120 50 65 115 64 80 115 20 ...
##  $ speed       : int  45 60 80 65 80 100 43 58 78 45 ...
##  $ type1       : chr  "grass" "grass" "grass" "fire" ...
##  $ is_legendary: int  0 0 0 0 0 0 0 0 0 0 ...
```

1. Types of variables

It is common to see different variables in a data set. There are many types of variables, but we can generally classify the variables as:

- Discrete/categorical: a variable that can only take values over a finite set of values (or levels), e.g., `is_legendary`, `name`
- Continuous: a variable that can take any numerical value over a continuous range, e.g., `wright_kg`, `height_m`
- Q: Are numerical variables always continuous? Can they be categorical?

2. Descriptive Statistics

Descriptive statistics are numerical summaries and plots used to describe and illustrate a data set. We will take a look at a few measures and tables commonly encountered in scientific journals.

2.1 Continuous variable

Some common measures to describe a continuous variable include:

- Mean, e.g., mean weight of pokemons is

```
mean(df$weight_kg, na.rm=T)
```

```
## [1] 61.3781
```

- Median, e.g., median weight of pokemons is

```
median(df$weight_kg, na.rm=T)
```

```
## [1] 27.3
```

- Variance/Standard deviation, e.g., variance of pokemons' weight is

```
var(df$weight_kg, na.rm=T)
```

```
## [1] 11958.46
```

2.1 Continuous variable (continued)

Some common measures to describe a continuous variable include:
- Minimum, e.g., maximum pokemon's weight is

```
min(df$weight_kg, na.rm=T)
```

```
## [1] 0.1
```

- Maximum, e.g., maximum pokemon's weight is

```
max(df$weight_kg, na.rm=T)
```

```
## [1] 999.9
```

- Range = Maximum - Minimum, and
- Interquartile range (IQR) = Q_3 - Q_1.

```
range(df$weight_kg, na.rm=T)
```

```
## [1]   0.1 999.9
```

2.2 Categorical variable

- The numerical summary of the continuous variable can be applied to categorical variable such as count.
- In many cases, the mode of a categorical and its distribution are more useful.
- The numerical summary of a categorical variable are usually summarized in a table.

```
table(df$type1)
```

```
   bug     dark   dragon electric    fairy fighting     fire   flying 
    72       29       27       39       18       28       52        3 
 ghost    grass   ground      ice   normal   poison  psychic     rock 
    27       78       32       23      105       32       53       45 
 steel    water 
    24      114 
```

2.2 Categorical variable (Continued)

The function `table()` can be used in a similar way to create contingency table.

```
table(df$type1, df$is_legendary)
```

```
           0   1
  bug       69   3
  dark      26   3
  dragon    20   7
  electric  34   5
  fairy     17   1
  fighting  28   0
  fire      47   5
  flying     2   1
  ghost     26   1
  grass     74   4
  ground    30   2
  ice       21   2
  normal   102   3
  poison    32   0
  psychic   36  17
  rock      41   4
  steel     18   6
  water    108   6
```

2.2 Categorical variable (Continued)

Sometimes it is more useful to report the proportions within the tables.

```
prop.table(table(df$type1))
```

```
        bug        dark      dragon    electric       fairy    fighting
0.089887640 0.036204744 0.033707865 0.048689139 0.022471910 0.034956305
       fire      flying       ghost       grass      ground         ice
0.064918851 0.003745318 0.033707865 0.097378277 0.039950062 0.028714107
     normal      poison     psychic        rock       steel       water
0.131086142 0.039950062 0.066167291 0.056179775 0.029962547 0.142322097
```

3. Hypothesis Test

Oftentimes, we are interested to investigate the relationships between multiple variables. For example,

- Is the mean weight of pokemons 50kg?
- Are legendary pokemons on average higher than the non-legendary pokemons?
- How does pokemons' attack and defense change after they grow?

These questions lead us to *hypothesis testing*.

3.1 Hypothesis testing: A crash course

- It begins with a question of interest that is similar to that of "a person charged with a crime".
- Statistical test acts as the "jury" for investigating the question of interest.
- The sample (i.e. data collected) is the "evidence".
- The statistical test is used to answer the question of interest based on the data collected.

3.2 Steps for hypothesis testing

1. Formulate the null and alternate hypothesis.
2. Choose and evaluate the appropriate test statistic (with R).
3. Assess the strength of the evidence against the null hypothesis.
4. Interpret the results.

Note that Step (2) is a tedious (and sometimes iterative) process. It cannot be checked off using "a few clicks".

3.2 Step 1: Null and alternate hypotheses

In hypotheses testing, we begin by translating a question of interest into the appropriate null and alternate hypotheses:

- Null hypothesis: Status quo statement that is commonly denoted as H_0. (An assertion that you want to prove wrong.)
- Alternate hypothesis: The answer the researcher is looking for, commonly denoted as H_1, H_A or H_a

3.2 Step 1: Null and alternate hypotheses

- Example: Is the average pokemon weight around 50kg?
  - H_0: Average pokemon weight is 50kg.
  - H_A: Average pokemon weight is not 50kg.
- Example: Is the average pokemon weight greater than 50kg?
  - H_0: Average pokemon weight is less than 50kg.
  - H_A: Average pokemon weight is greater than 50kg.

The first example is called a "two-sided test", the second is called a "one-sided test".

3.2 Step 2: Statistical tests

Step 2 involves choosing the appropriate test statistics. In this workshop, we will briefly discuss several common statistical tests:

- t-Test
- one-sample t-test
- two-sample t-test
- paired t-test
- Analysis of Variance (ANOVA)
- z/Wald-test

3.2 Step 3: Strength of evidence

- In most statistical tests, a p-value will be produced.
- The p-value is the probability of finding results equal or more extreme than the observed results (data), given that the null hypothesis (H_0) is true.
- The smaller the p-value, the more evidence we have against the null hypothesis.
- The default significance levels are 0.01, 0.05 and 0.10.
  - When the p-value is less than the significance level (of your choice), we say that we have evidence against the null hypothesis in favor of the alternate hypothesis.
  - When the p-value is greater than the default value, we say that we do not have sufficient evidence against the null hypothesis. Sometimes, we say "we do not reject the null hypothesis".
  - However, we almost always avoid saying "we accept the null hypothesis".

3.2 Drawing conclusion

- The final step in hypothesis testing is to draw conclusion in the words of the problem.
- Example: Is the average pokemon weight around 50kg?
  - H_0: Average pokemon weight is 50kg.
  - H_A: Average pokemon weight is not 50kg.
- If the p-value is less than 0.05, we say that there is evidence against H_0, in favour of H_A, i.e. the data suggests that the average pokemons' weight is not 50kg.
- If the p-value is greater than 0.05, we say that there is not enough evidence against H_0, i.e. the data suggests that the average pokemons' weight is around 50kg.

3.2 Drawing conclusion

- Example: Is the average pokemon weight greater than 50kg?
  - H_0: Average pokemon weight is less than 50kg.
  - H_A: Average pokemon weight is greater than 50kg.
- Q: What can we conclude if the p-value is less than 0.05?

4. T-Tests

- One of the most commonly used statistical test is the t-test.
- The **one-sample t-test** is used to compare the mean of a variable to a hypothetical value. In most cases, the hypothetical value comes from theory.
- The **two-samples t-test** is used to compare the means of two variables. It is often used to determine whether a treatment has an effect on the population of interest, or whether two groups are different from one another.
- The **paired t-test** is commonly used to investigate the difference of a variable pre- and post-treatment. Oftentimes, every subject of the study produces a pair of observations, or two similar subjects will be paired up.

4.1 One-sample t-test

- For a one-sample t-test, the assumptions are:

1. The population of the variable from which the sample is drawn from is **independently identically distributed** (IID) from Normal distribution.
2. If the sample size is large enough, the assumption in (1) is not necessary because Central Limit Theorem applies.

4.1 Example: Is the average pokemon weight around 50kg?

- H_0: Average pokemon weight is 50kg.
- H_A: Average pokemon weight is not 50kg.

```
t.test(df$weight_kg, mu=50)
```

```
    One Sample t-test

data:  df$weight_kg
t = 2.9078, df = 780, p-value = 0.003744
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 53.69681 69.05940
sample estimates:
mean of x
  61.3781
```

- Q: What can you say from this?

4.1 Example: Is the average pokemon weight around 50kg?

- The p-value is less than 0.05 and hence, we have evidence against H_0, i.e. the average pokemons' weight is not 50.
- Q: But is it higher or lower?

```
t.test(df$weight_kg, alternative="greater", mu=50)
```

```
    One Sample t-test

data:  df$weight_kg
t = 2.9078, df = 780, p-value = 0.001872
alternative hypothesis: true mean is greater than 50
95 percent confidence interval:
 54.9341     Inf
sample estimates:
mean of x
  61.3781
```

4.2 Two-samples t-test

- Pokemons can be categorized as legendary and non-legendary. Suppose we want to compare the weight/height/hp/... between these two groups.
- As in the one-sample t-test, there are assumptions for using the two-samples t-test:

1. The population in which each sample was drawn from is **independently normally distributed**.
2. The population variances are similar.
   - This is not as important when using R because the default in R is to assume unequal population variances. The results of tests for equal and unequal population variances will be the same if the population variances are the same.
3. When sample size is large enough, (1) is not important since Central Limit Theorem applies.

4.2 Example: Is avg weight of legendary pokemons the same as the non-legendary?

- H_0: The avg weight of legendary and non-legendary pokemons are the same.
- H_A: The avg weight of legendary and non-legendary pokemons are NOT the same.

```
t.test(df$weight_kg[df$is_legendary == 1],
       df$weight_kg[df$is_legendary == 0])
```

```
    Welch Two Sample t-test

data:  df$weight_kg[df$is_legendary == 1] and df$weight_kg[df$is_legendary == 0]
t = 5.1176, df = 69.188, p-value = 2.65e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  92.34922 210.33790
sample estimates:
mean of x mean of y
199.35072  48.00716
```

- Q: What can you say from this?

4.2 Example: comparing weights of pokemons

- H_0: The avg weight of legendary and non-legendary pokemons are the same.
- H_A: The avg weight of legendary and non-legendary pokemons are NOT the same.

```
t.test(df$weight_kg[df$is_legendary == 1],
       df$weight_kg[df$is_legendary == 0])
```

```
    Welch Two Sample t-test

data:  df$weight_kg[df$is_legendary == 1] and df$weight_kg[df$is_legendary == 0]
t = 5.1176, df = 69.188, p-value = 2.65e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  92.34922 210.33790
sample estimates:
mean of x mean of y
199.35072  48.00716
```

- Q: What can you say from this?

4.2 Example: comparing heights of pokemons

- H_0: The avg height of legendary pokemons are greater than that of non-legendary pokemons.
- H_A: The avg height of legendary pokemons are lower than that of non-legendary pokemons.

```
t.test(df$height_m[df$is_legendary == 1],
       df$height_m[df$is_legendary == 0],
       alternative = "less")
```

```
    Welch Two Sample t-test

data:  df$height_m[df$is_legendary == 1] and df$height_m[df$is_legendary == 0]
t = 5.6684, df = 71.593, p-value = 1
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
    -Inf 1.58583
sample estimates:
mean of x mean of y
 2.281159  1.055618
```

- Q: What can you say from this?

4.3 Paired t-test

- When the two samples are highly correlated, assumptions of two-sample t-test are violated.
- E.g., pokemons' attack and defense before and after they evolve (supergrow) are highly correlated because different pokemon can have difference attack and defense abilities.
- Q: Are pokemons' attack on average unchanged after they evolve?
- Q: Are pokemons' defense on average higher after they evolve?

4.3 Example: comparing attacks of pokemons

- *H*0: Pokemons' attacks are the same before and after they evolve.

```
t.test(df$attack, df$sp_attack, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  df$attack and df$sp_attack
## t = 5.1137, df = 800, p-value = 3.956e-07
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  4.036861 9.066759
## sample estimates:
## mean difference
##         6.55181
```

- Q: What can you say from this?

4.3 Example: comparing defenses of pokemons

```r
t.test(df$defense, df$sp_defense,
       alternative = "less", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  df$defense and df$sp_defense
## t = 2.0698, df = 800, p-value = 0.9806
## alternative hypothesis: true mean difference is less than 0
## 95 percent confidence interval:
##      -Inf 3.766043
## sample estimates:
## mean difference
##        2.097378
```

- Q: What is the null hypothesis of this test?
- Q: What can you say about this?

5. Linear Regression

The linear regression is also known as linear model. It is widely used in data analysis because:

- the model assumptions are often found satisfactory among many data sets; and
- the interpretation of each parameter in the model is easy and clear.

When the assumptions of the linear regression model are satisfied, the model is powerful in terms of inference and interpretation.

5.1 Explanatory and response variables

The most common goal in research is to understand relationship between variables. These variables are typically categorized as:

- Response variable (or dependent variable): the hp of a pokemon
- Explanatory variable (or independent variable): A measure in the study used to explain, predict or influence the response variable.

In this workshop, we will only consider response variables that are continuous.

- For non-continuous response, you need to consider Generalized Linear Regression models, e.g.,
  - Logistic Regression for binary response
  - Poisson Regression for count response

5.2 Model assumptions

A simple linear regression model assumes that:

- given the predictors, the expectation of the response is a linear function.
- the errors are normally distributed.
- the errors are independent of one another.
- the errors have mean zero and equal variance.

5.3 Simple Linear Model

- A simple linear model investigates possible linear relationship between two random variables.
- The response variable is a continuous variable.
- The explanatory variable can be of any type.
- Suppose we are interested in the association of pokemons' `hp` with their weight and height.

```
hp_lm = lm(hp ~ weight_kg + height_m, data = df)
```

5.4 z/Wald-test

- Use to test the statistical significance of association between the response variable and an explanatory variable.
- We say an explanatory variable is *statistically significant* if it has strong association with the response variable.
  - $H0$: The explanatory variable is NOT statistically significant for predicting the response variable.
  - $H1$: The explanatory variable is statistically significant for predicting the response variable.
- Q: Are the weight and height statistically significant?

5.4 z/Wald-test

- We use z/Wald test to examine the statistical significance of an explanatory variable.

```
summary(hp_lm)
```

```
##
## Call:
## lm(formula = hp ~ weight_kg + height_m, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -106.941  -13.161   -2.085    9.317  183.731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 56.06663    1.21586  46.113  < 2e-16 ***
## weight_kg    0.05131    0.00966   5.311 1.42e-07 ***
## height_m     8.53407    0.97787   8.727  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 
```

6. Analysis of variance (ANOVA)

- The name analysis of variance can be misleading. It is actually a test on means over more than 2 samples.
- Remember that we have t-test to test on means of 1 or 2 samples.
- Pokemons have different types:

```
## tp
##    bug   fire  grass normal  other  water
##     72     52     78    105    380    114
```

- We can use One-way ANOVA to test whether different types of pokemons have similar weight/height/hp/attack/defense ...

6.1 One-way ANOVA

In one-way ANOVA, we test the null hypothesis that k populations all have the same mean

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

against the alternative hypothesis that the population means are not all equal.

The assumptions of one-way ANOVA are the same as those of the pooled-variance two-samples t-test:

1. The samples are independent simple random samples from the populations.
2. The populations are normally distributed.
3. The population variances are equal. ANOVA works poorly if the variances are extremely different.

Example 6.1: Pokemons' hp over different types

- $H0$: Different types of pokemons have the same level of hp.

```
aov_model = aov(hp~factor(type1), data=df)
summary(aov_model)
```

```
                Df Sum Sq Mean Sq F value   Pr(>F)
factor(type1)   17  28555  1679.7   2.452 0.000936 ***
Residuals      783 536473   685.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

- Q: What can you say from this?

6.2 Two-way ANOVA

- In reality, we often want to understand the impact of two independent variables and their combination on the response variable.
- We can compare the hp/weight/height of pokemons that are of different types and are either legendary or non-legendary.
- Instead of performing two individual t-tests or one-way ANOVA tests, we could perform a two-way ANOVA.
- In general, when the response variables of the one-way ANOVA are the same, we try to use one model.
- As mentioned earlier, performing multiple statistical tests is not only tedious, but also increases the risk of a Type I error.

7. Final Remarks

- So far we have introduced several tests and ways to examine statistical significance.
- But we cannot cover everything that you need to be aware of when you start to perform a formal statistical analysis.
- We would like to emphasize some key caveats that would help you avoid drawing incorrect conclusions from your analysis results.

7.1 Caveat 1: Statistical significance vs practical significance

- Statistical inference techniques test for statistical significance.
- Statistical significance means that the effect observed in a sample is very unlikely to occur if the null hypothesis is true.
- Whether this observed effect has practical importance is an entirely different question. The experts in the field of interest determine whether these results have any practical importance.

7.2 Caveat 2: Danger of over reliance on p-values

The ASA's Statement on p-values:

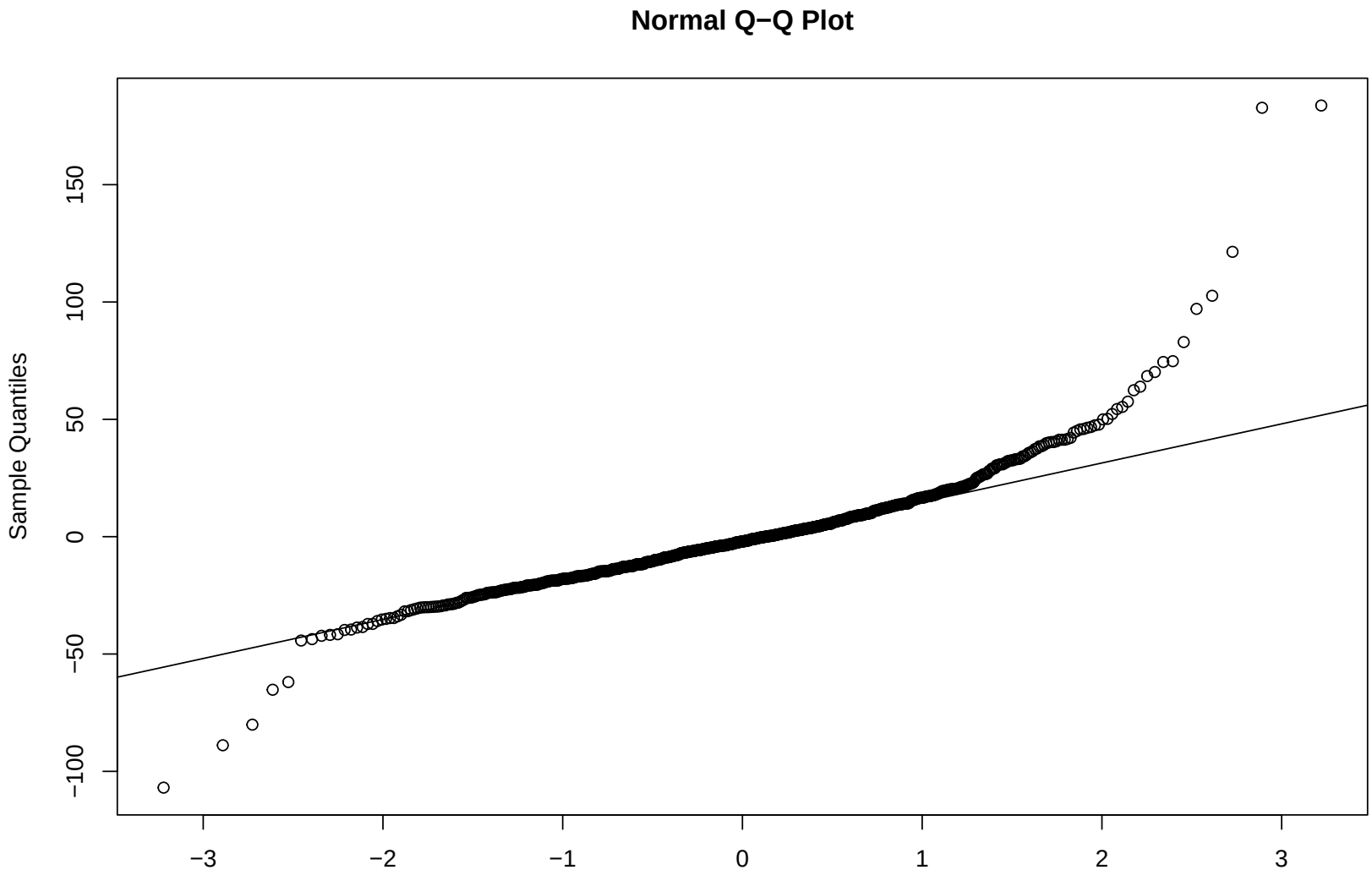
- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

7.3 Caveat 3: Model diagnostics

- The statistical test results are only valid when the test assumptions are satisfied by the data.
- We did not check model assumptions when we performed the tests above, but this is a step that we cannot skip when doing actual analysis.
- There are a variety of model diagnostics test for different model assumptions.
  - **quantile-quantile plot** (QQ-plot). When the assumptions of residuals normality is met, we expect the points to lie on a straight line.
  - **residuals against the explanatory variables**. When the assumption of independent error is met, we expect the points to scatter randomly around the the horizontal line $y = 0$.

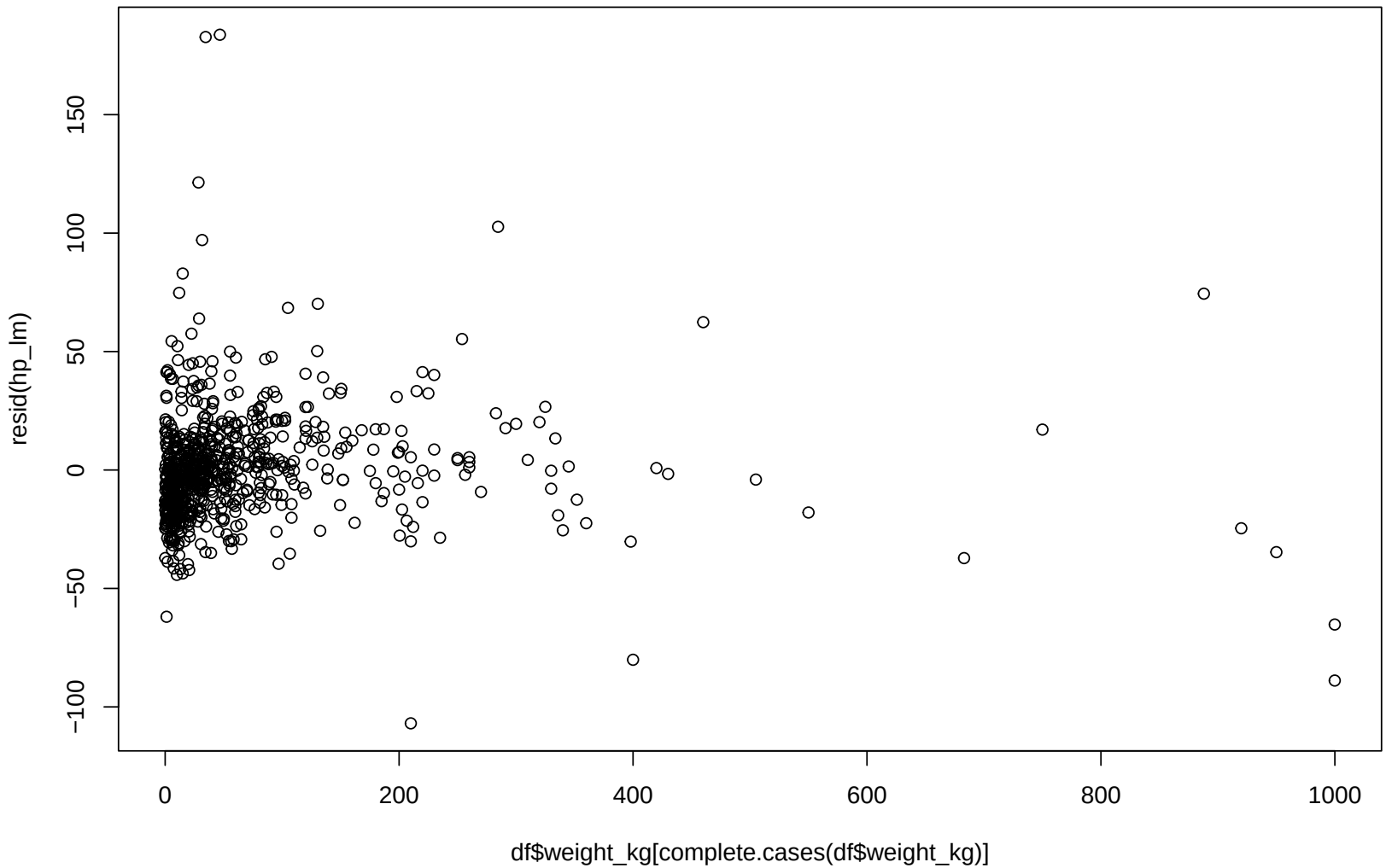
An example of the QQ-plot

```
qqnorm(resid(hp_lm))
qqline(resid(hp_lm))
```



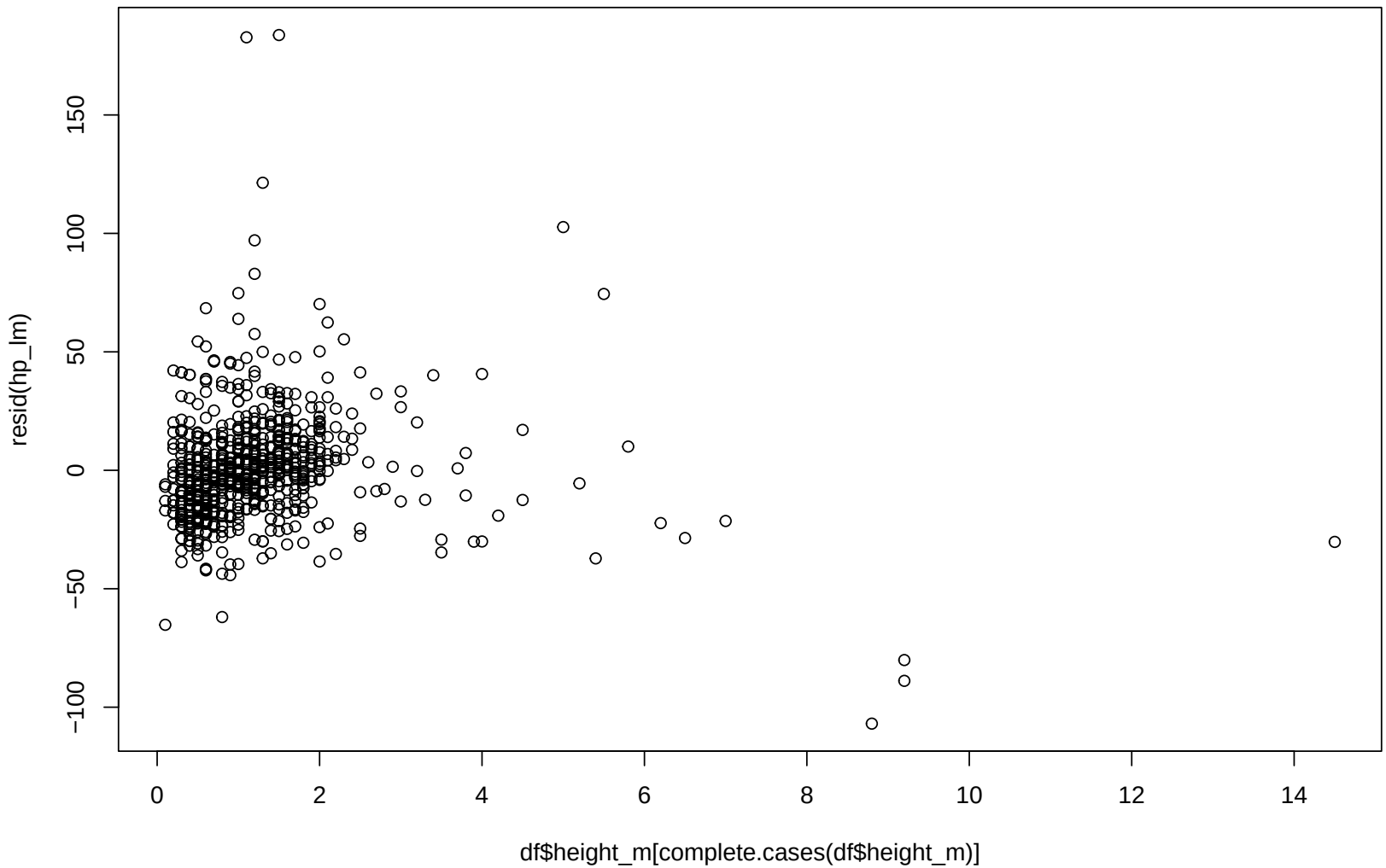
An example of a residual plot

```
plot(df$weight_kg[complete.cases(df$weight_kg)], resid(hp_lm))
```



An example of a residual plot

```
plot(df$height_m[complete.cases(df$height_m)], resid(hp_lm)
```



7.4 Caveat 4: Model fitting procedure

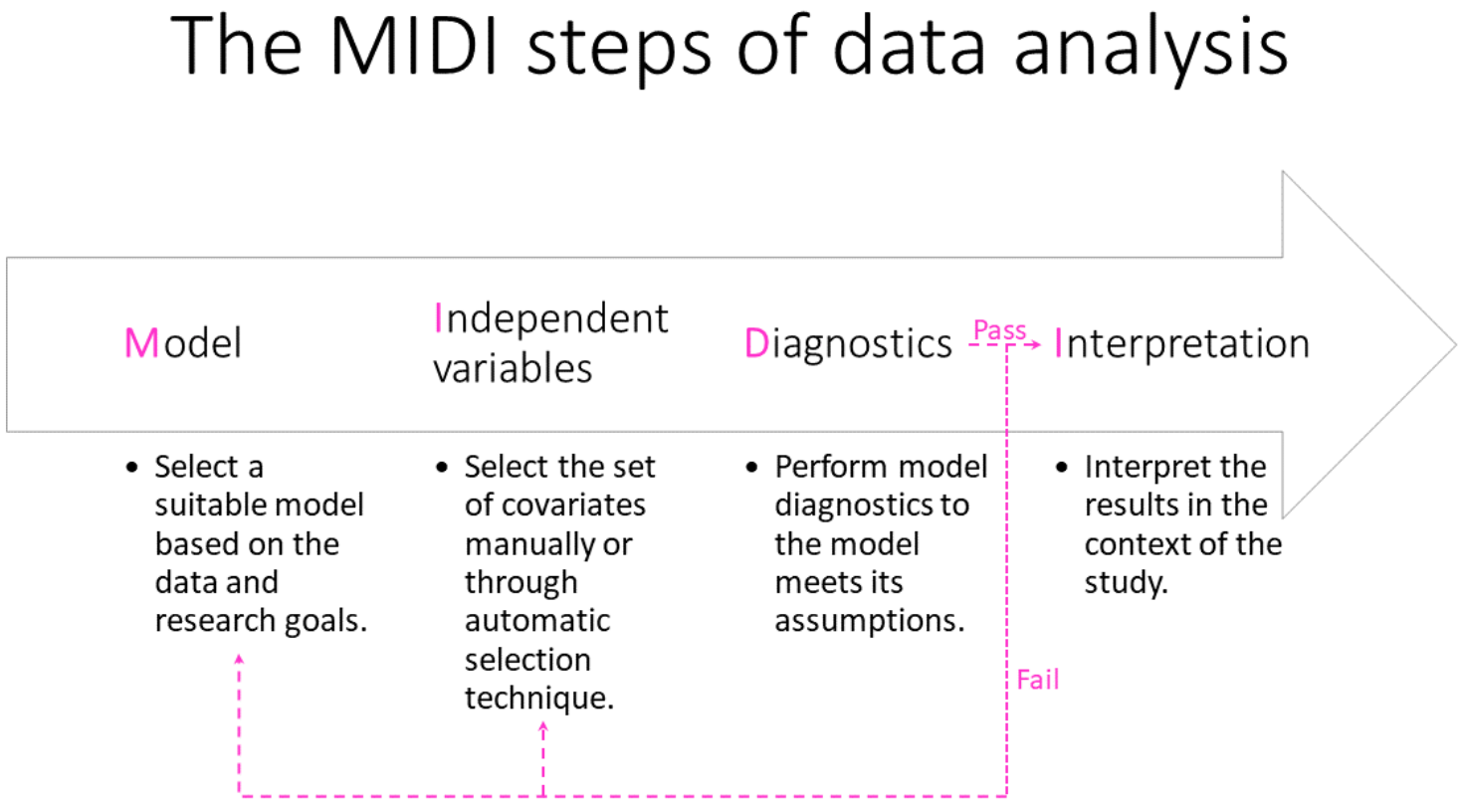


Figure 1: Recommended steps to data analysis

8. Practice Questions

- Here we prepare some practice questions for self examination and for fun.
- Recall that there are multiple types of pokemons

```
table(df$type1)
```

```
##
##      bug     dark   dragon electric     fairy fighting     fire   flying
##       72       29       27       39       18       28       52        3
##    ghost    grass   ground      ice   normal   poison  psychic     rock
##       27       78       32       23      105       32       53       45
##    steel    water
##       24      114
```

8.1 Ex: Compare pokemons of different types

- Q1: What are the weights of fire pokemons compared to that of grass pokemons?
- Q2: What are the attacks of poison pokemons compared to that of dark pokemons?
- Q3: Are fighting pokemons on average taller than the psychic pokemons?

8.2 Ex: Compare Pokemons before/after they "evolve"

- After evolving, a pokemon can perform a super-attack (`sp_attack`) or a super-defense (`sp_defense`).
- Q1: Are pokemons' attack increased more than their defense after evolving?
- Q2: Is the increase of attack of legendary pokemons' higher than that of the non-legendary pokemons' after they evolve?

8.3 Ex: What kind of pokemon is likely heavier?

- Say we are interested in the weight of pokemons, because we can simply hold a cute Pickachu in arms as if it were a kitten, but certainly not a fully evolved Charizard!

Our advice

- Visualize your data with suitable graphs.
- Check the model assumptions.
- If the normality assumption is not reasonable, there are other options available.
- Avoid extrapolating.
- Correlation does not imply causation.
- Every model has its strengths and limitations. When in doubt, get help. The SCSRU offers free 1-1 consultation to all UWaterloo researchers.

"All models are wrong, but some are useful." — George Box

Next steps

Now that you have reviewed how to perform some common statistics with R, you can learn more about the different models, or explore other models such as

- model diagnositics
- exploratory data analysis,
- linear regression with R,
- generalized linear models,
- count data analysis, etc.

The SCSRU organize similar workshops to this on a regular basis to improve quality of research and data literacy among the UWaterloo community. We also provide 1-1 free consultation to all researchers on campus. More information are available on our website.

Thank you!

The Statistical Consulting and Survey Research Unit (SCSRU) is the unit through which the Department of Statistics and Actuarial Science provides statistical advice to those working on research problems.