Introductory Statistics with R

SCSRU Workshop

Statistical Consulting and Survey Research Unit
University of Waterloo

2023-09-16

Agenda

- Type of Variables
- Descriptive Statistics
- Outliers and Extreme Values
- Hypothesis Test: A Gentle Introduction
  - t-Test
  - ANOVA Test
  - Correlation Test
- Linear Regression
  - Simple Linear Regression & Multiple Linear Regression
  - Model Diagnostics
- Modelling Fitting Procedure

A quick review

Throughout this workshop, we will use 2 data sets: `caliRain.csv` and `drinks.csv` as motivating examples.

1. Did you receive 2 CSV files a few days ago? Please save both files in the same folder for the purpose of this workshop.
2. Do you remember where you saved the file? Please set the working directory to the folder those files were saved.

```
setwd("Your_Working_Directory")
```

3. Please import both data set into the current R environment.

```
drinks_df <- read.csv("drinks.csv")
rain_df <- read.csv("caliRain.csv")
```

A look at the caliRain.csv data

```
head(rain_df)
```

	STATION	PRECIP	ALTITUDE	LATITUDE	DISTANCE	SHADOW
1	Eureka	39.57	43	40.8	1	1
2	RedBluff	23.27	341	40.2	97	2
3	Thermal	18.20	4152	33.8	70	2
4	FortBragg	37.48	74	39.4	1	1
5	SodaSprings	49.26	6752	39.3	150	1
6	SanFrancisco	21.82	52	37.8	5	1

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
- Continuous

1.1 Continuous variables

A continuous variable is a variable that can take any value over a continuous range. For example,

- age in years,
- number of work hours,
- midterm scores, etc.

In the `caliRain.csv` data set, the variable `PRECIP`, `ALTITUDE`, `LATITUDE` and `DISTANCE` are considered continuous variable.

1.2 Discrete variables

Discrete variables are sometimes known as categorical variables or qualitative variables. A categorical variable is a variable that can only take values over a finite set of values (or levels). Examples include:

- A university student's major.
- A person's blood type.
- The entrees on a menu.
- A person's eye colour.
- A person's level of agreement about a statement.

Binary variable: categorical variables with only 2 levels.

In the `caliRain.csv` data set, the variable `SHADOW` can only take two values: 1 and 2. We call `SHADOW` a binary variable.

1.2.2 Nominal variables

A categorical variable with no specific ordering is also called a nominal variable. Examples include:

- A university student's major.
- A person's blood type.
- The entrees on a menu.
- A person's eye color.

1.2.3 Ordinal variables

A categorical variable with natural ordering is also called an ordinal variable. Examples include

- A person's eye color.
- A person's level of agreement about a statement.

Notice that the example of "a person's eye color" shows up as nominal and ordinal variable. Why?

Is the variable type fixed?

We cannot determine the variable type by its name. To accurately describe a variable, we need to consider how it is recorded.

A common example is age. In some studies, age is recorded as an exact value, e.g. 25, 35.5, 80, etc. This age variable is considered a continuous variable. Other studies may require respondents to select the category in which their age falls in, e.g. $<$20, 21-25, 80$+$ etc. This age variable is a categorical variable.

There are variables that are neither continuous nor discrete. An example is text responses.

Practice 1.1

There are 6 variables in `drinks.csv`, namely

- drink
- hasMilk
- temp
- fat
- carb
- calories

What type of variables are they?

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,
- Median (or Q_2),
- Variance/Standard deviation,
- Minimum,
- Maximum,
- Range = Maximum - Minimum, and
- Interquartile range (IQR) = Q_3 - Q_1.

In the presence of extreme values, median and IQR are preferable.

Practice 2.1

- Can you find the mean, median and variance of `PRECIP` in the `rain_df`?
- Have you heard about the function `summary()` and `fivenum()`?

```
summary(rain_df$PRECIP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.660   9.565  15.345  19.807  21.198  74.870
```

```
fivenum(rain_df$PRECIP)
```

```
## [1]  1.660  9.440 15.345 21.820 74.870
```

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.

```
milk_table<- table(drinks_df$hasMilk)
milk_table
```

```
  Milk  Nonmilk
    30        6
```

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

```
table_of_milk_by_temp<- table(drinks_df$hasMilk,
                              drinks_df$temp)
table_of_milk_by_temp
```

```
          Cold Hot
  Milk      20  10
  Nonmilk    6   0
```

This table is known as a 2x2 contingency table.

Sometimes it is more useful to report the proportions within the tables. There are two ways to do so.

```
prop.table(milk_table)
```

```
    Milk    Nonmilk
0.8333333 0.1666667
```

```
proportions(milk_table)
```

```
    Milk    Nonmilk
0.8333333 0.1666667
```

To convert the proportions within the table into percentages, and round the percentages to 2 decimal places.

```
round(100*prop.table(milk_table),2)
```

```
  Milk  Nonmilk
 83.33   16.67
```

Note that, scientific journals that follow the APA formatting require the percentages to be reported as whole numbers, i.e. no decimal place.

Practice 2.2

Recall the 2x2 contingency table `table_of_milk_by_temp`, try using the `prop.table()` or `proportions()` function on the contingency table. What are these proportions about?

Practice 2.3

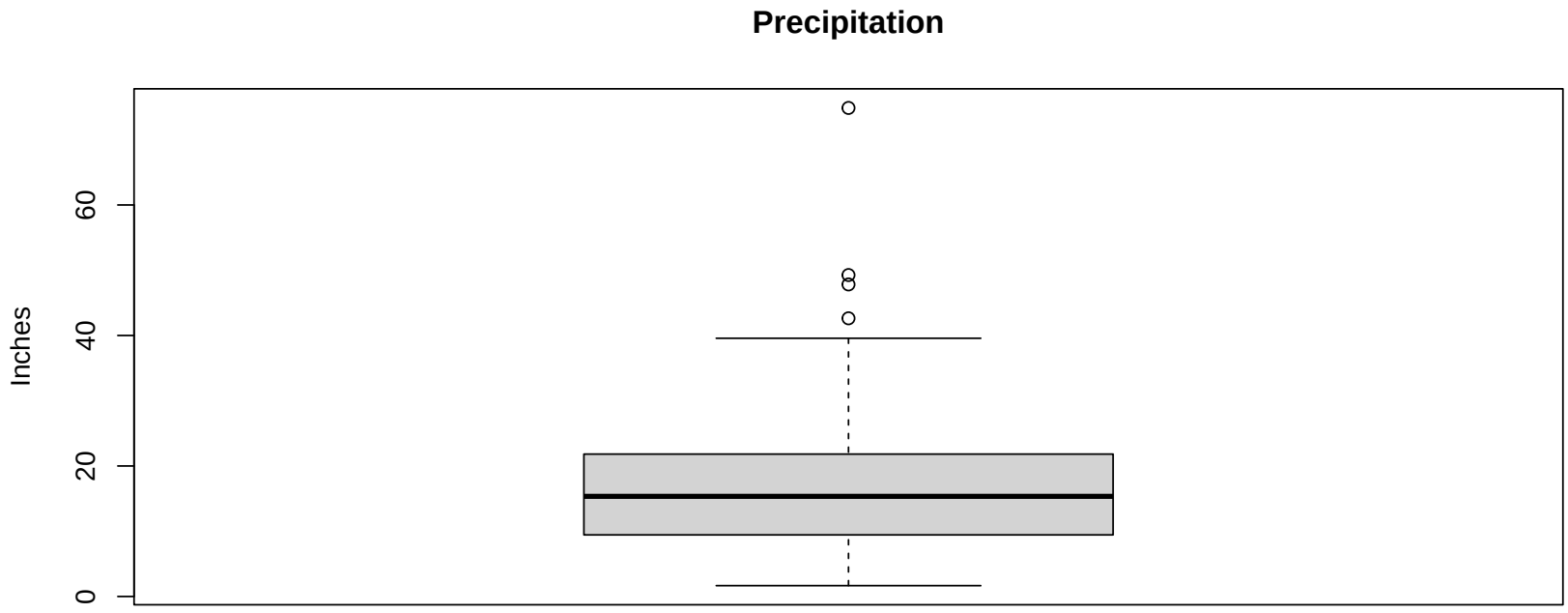
- Both `prop.table()` or `proportions()` functions have an argument called `margin` that takes the value 1 or 2.
- When the margin is specified, the outputs are conditional proportions.
- Notice that when `margin = 1`, the row sums to 1, whereas when `margin = 2`, the columns sum to 1.

Find the proportion of cold drinks that contain milk.

3. Outliers and extreme values

- The boxplot is often used to visualize the distribution of a numeric variable, and potential outliers.
- The oultiers are presented as dots or points beyond the box and its whiskers.
- The rule used to identify the outliers is called the $1.5 \times IQR$ rule, where $IQR = Q_3 - Q_1$.

```r
boxplot(rain_df$PRECIP, main="Precipitation",
        ylab= "Inches")
```



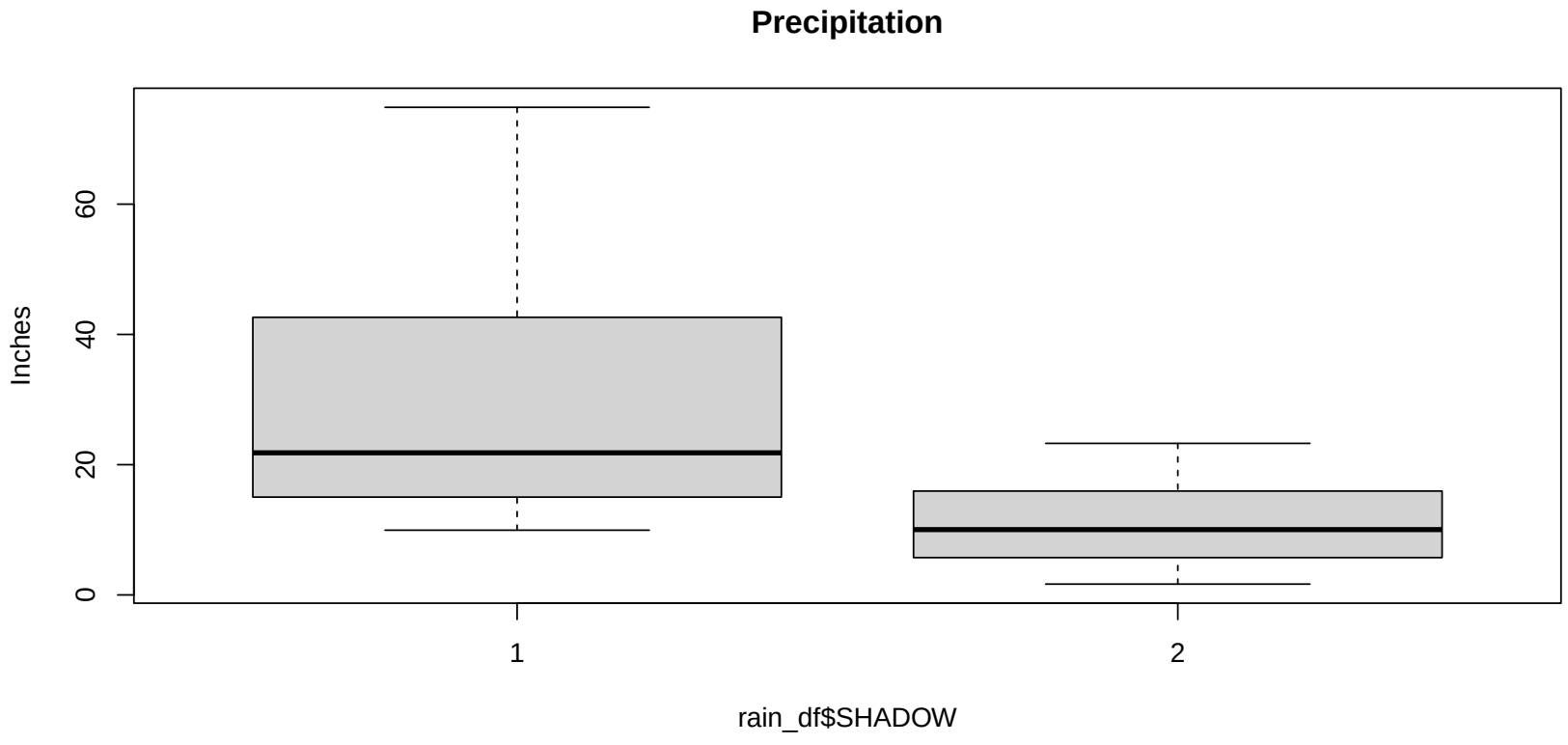
Why 1.5?

- Tukey who introduced the $1.5 \times IQR$ rule claimed that, "2 was too big, and 1 was too small."
- This rule suggests whether a value is potentially an outlier.
- However, experts opinion is crucial when deciding whether an observation is an outlier, or an extreme value.
- An extreme value may contain interesting information and must not be dismissed without careful thoughts.

Side-by-side boxplots

In the side-by-side boxplots, notice that there is no potential outlier. What happened to the circles in the first boxplot?

```
boxplot(rain_df$PRECIP~rain_df$SHADOW, main="Precipitation",
        ylab= "Inches")
```



3. Exploratory data analysis

Exploring the data through descriptive summaries and graphical tools is an essential step. We can better understand

- the structure of the data,
- the variables and their distributions, and
- existing extreme values and outliers.

Although this step is tedious, it prepares the data and the analyst for more sophisticated analysis. Some may refer to this as exploratory data analysis.

4. Introduction to Hypothesis Test

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

- Does the distance from Pacific Ocean affect precipitation?
- Is the midterm average this term higher than the average last term?
- Is my average grocery purchase every week around $50?

The questions lead us to *hypothesis testing*.

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.

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".

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

Example 4.1: Is my average grocery purchase every week around \$50?

- H_0: My average grocery purchase is \$50.
- H_A: My average grocery purchase is not \$50.

Examples

Example 4.2: Is the midterm average this term higher than the average last term?

- H_0: The midterm average this term is the same as the average last term.
- H_A: The midterm average this term is higher than the average last term.

Example 4.3: Does the distance from Pacific Ocean affect precipitation?

- H_0: Precipitation is not affected by the distance from Pacific Ocean.
- H_A: Precipitation is affected by the distance from Pacific Ocean.

Example 4.1 and 4.3 are called "two-sided test". Example 4.2 is called a "one-sided test".

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
- Analysis of Variance (ANOVA)
- Pearson Correlation Test
- Linear Regression

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".

Drawing conclusion

The final step in hypothesis testing is to draw conclusion in the words of the problem.

Example 4.1: Is my average grocery purchase every week around \$50?

- H_0: My average grocery purchase is \$50.
- H_A: My average grocery purchase is not \$50.

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 my true average grocery purchase is not \$50.

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 my true average grocery purchase is \$50.

Example 4.2

Is the midterm average this term higher than the average last term?

- H_0: The midterm average this term is the same as the average last term.
- H_A: The midterm average this term is higher than the average last term.

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 midterm average this term is higher than the average last term.

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 the midterm average this term is the same as (or similar to) the average last term.

4.1 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.

4.2 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.

5. 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.

5.1 One-sample t-test

Have you ever asked:

- Is my coffee purchase around $50 per week?
- Is my grade higher than the class average?
- Is the average annual precipitation in California around 25 inches?

The value for comparison (i.e. $50, class average, 25 inches) are known as hypothetical values. It is a value that is obtained through literature that acts as a point of comparison.

Assumptions

It is important to check the assumptions of the tests before conducting a statistical test. The results from a hypothesis test may not be valid if any of the assumption was violated.

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.

Example 5.1

H_0 : The average annual precipitation in California is 25 inches. H_A : The average annual precipitation in California is not 25 inches.

We can perform this two-sided one-sample t-test in R:

```
t.test(rain_df$PRECIP, mu=25)
```

```
    One Sample t-test

data:  rain_df$PRECIP
t = -1.7112, df = 29, p-value = 0.09773
alternative hypothesis: true mean is not equal to 25
95 percent confidence interval:
 13.60088 26.01379
sample estimates:
mean of x
 19.80733
```

The p-value is greater than 0.05 and hence, we claim no evidence against H_0, i.e. the average annual precipitation in California is 25 inches.

The average annual precipitation is 19.81 inches. How is this 25 inches?

5.2 Two-samples t-test

The more common questions we often need to answer require us to compare two groups (samples). For example,

- Are two marketing campaigns equally effective?
- Do males and females have a different mean body mass index?
- Do precipitation differs due to shadow?

Another way to think about this is that we are interested to investigate the relationship between a continuous and a categorical variable.

Assumptions

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.

Example 5.2

Is the average annual precipitation affected by `SHADOW`?

- H_0: The precipitation is not affected by `SHADOW`.
- H_A: The precipitation is affected by `SHADOW`.

In another words,

- H_0: The precipitation on the `Leeward` side is the same as the precipitation on the `Westward` side.
- H_A: The precipitation on the `Leeward` side is not the same as the precipitation on the `Westward` side.

Example 5.2: R codes and output

```
t.test(rain_df$PRECIP~rain_df$SHADOW)
```

```
    Welch Two Sample t-test

data:  rain_df$PRECIP by rain_df$SHADOW
t = 3.5309, df = 14.01, p-value = 0.003321
alternative hypothesis: true difference in means between group 1
95 percent confidence interval:
  7.743558 31.702957
sample estimates:
mean in group 1 mean in group 2
       30.98385        11.26059
```

Example 5.3

Is there higher precipitation on the `Westward` side?

- H_0: The precipitation on the `Leeward` side is the same as the precipitation on the `Westward` side.
- H_A: The precipitation on the `Leeward` side is less than the precipitation on the `Westward` side.

or

- H_A: The precipitation on the `Westward` side is more than the precipitation on the `Leeward` side.

- In R, the default test is the two-sided test.
- Example 5.3 calls for a one-sided test.
- To perform a one-sided test, we need to make changes to the `alternative` in the code.
- If the alternate hypothesis is $\mu_{Group1} < \mu_{Group2}$, then we have `alternative ="less"`. Otherwise, set `alternative ="greater"`.
- By default, *Group*1 is the category in which the category's name comes first in alphabetical order.

R codes and output

```
t.test(rain_df$PRECIP~rain_df$SHADOW, alternative = "less")
```

```
    Welch Two Sample t-test

data:  rain_df$PRECIP by rain_df$SHADOW
t = 3.5309, df = 14.01, p-value = 0.9983
alternative hypothesis: true difference in means between group 1
95 percent confidence interval:
     -Inf 29.56122
sample estimates:
mean in group 1 mean in group 2
       30.98385        11.26059
```

Practie 5.1

Suppose the experts want to find out whether `LATITUDE` is affected by `SHADOW`. Choose an appropriate test to help the experts.

6. 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): An outcome of the study or of interest.
- 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.

7. Analysis of variance (ANOVA)

- The name analysis of variance can be misleading. It is actually a test on means.
- In one-way ANOVA, the pooled variance two sample t-test is extended to more than two samples.

Example 8.1: Consider the `oats` data set from the `MASS` library

```
library(MASS)
head(oats)
```

```
  B             V      N   Y
1 I      Victory 0.0cwt 111
2 I      Victory 0.2cwt 130
3 I      Victory 0.4cwt 157
4 I      Victory 0.6cwt 174
5 I Golden.rain 0.0cwt 117
6 I Golden.rain 0.2cwt 114
```

Example 7.1

Briefly, the variables in this data set are:

- B: Blocks
- V: Varieties
- N: Nitrogen treatment
- Y: Yield of the crop

Suppose we are interested to find out whether the average yield (response variable) of each variety (within the independent variable, V) are the same.

There are 3 varieties of oats:

```
levels(oats$V)
```

```
[1] "Golden.rain" "Marvellous"  "Victory"
```

Example 7.1

It is tempting to compare the average yield of the varieties in pairs using t-test:

- `Golden.rain` vs `Marvellous`
- `Golden.rain` vs `Victory`
- `Marvellous` vs `Victory`

However, such paired comparisons have limitations:

- the process can be tedious when there are many pairs
- the risk of a type I error increases when making multiple statistical tests. Type I error means rejecting the null hypothesis when it's actually true.

7.1 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 7.1: Analysis

```
oats_aov <- aov(Y~V,data=oats)
summary(oats_aov)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
V            2   1786   893.2   1.228  0.299
Residuals   69  50200   727.5
```

- Before making any interpretation, we must check the QQ-plot of the residuals to ensure that the model meets its assumptions. We will cover model diagnostics briefly later.
- Assuming the model fit is good, the corresponding p-value is 0.299, indicating no sufficient evidence that the three varieties of oats have different yields.

7.2 Two-way ANOVA

- In reality, we often want to understand the impact of two independent variables and their combination on the response variable.

Example 7.2: Recall the `drinks_df` data set. Suppose we are interest to find out whether milk content (`hasMilk`) and temperature (`temp`) can affect the amount of calories.

- Instead of performing two individual t-tests, we would 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.

Example 7.2: Analysis

```
drinks_aov <- aov(calories~factor(hasMilk) + temp,
                  data=drinks_df)
summary(drinks_aov)
```

```
                 Df Sum Sq Mean Sq F value  Pr(>F)
factor(hasMilk)   1 191753  191753  24.893 1.9e-05 ***
temp              1    807     807   0.105   0.748
Residuals        33 254197    7703
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The variable `hasMilk` has corresponding p-value smaller than 0.05. This indicates strong evidence that the groups within that variable have different averages of calories.
- The variable `temp` has corresponding p-value larger than 0.05. This indicates that holding everything else constant, there is no evidence the `Hot` and `Cold` drinks have different calories on average.

- A common practice is to remove the insignificant variable from the model. However, we advice relying on significance entirely.
- Every variable contributes differently to a model.
- Some variable contribute by explaining the variability of the response variable. These variables will appear with small p-values.
- Some variable contirbute by holding the structure of a model. They may not be significant, but they help the model meets its assumptions.
- It is important to check the model fit when variable(s) is/are added or removed from the model.

7.3 Post-hoc Tests

If there is strong evidence that not all the population means are equal for one variable, the next question is which categories are different. The ANOVA does not tell us which population means are different.

To explore where the difference lies, we perform the post-hoc tests. The post-hoc tests control for family-wise error rate. Here are a few common post-hoc tests:

- Fisher's Least Significant Difference (LSD), `LSD.test()`,
- Bonferroni correction, `pairwise.t.test(x, g, p.adjust.method="bonferroni")`,
- Tukey's Honest Significant Different, `TukeyHSD()`, and
- Scheffe's, `ScheffeTest()`.

8.1 Correlation

- Pearson correlation, r ,is used to measure the linear relationship between two continuous variables.
- It is also unitless.
- Correlation is between -1 and 1.
  - If $r \approx 0$, the linear relationship between two variables is weak.
  - If $r \approx 1$, there is a strong positive linear relationship between two variables.
  - If $r \approx -1$, there is a strong negative linear relationship between two variables.

Example 8.1

Suppose we are interested to evaluate the correlation between ALTITUDE and PRECIP in R:

```
cor(rain_df$PRECIP, rain_df$ALTITUDE)
```

```
[1] 0.3020067
```

- It does not matter which variable is first because the correlation between X and Y is the same as the correlation between Y and X.

8.2 Pearson's correlation test

The goal of this hypothesis test is to test the null hypothesis that the true correlation is equal to zero.

$$H_0 : r = 0$$

If p-value < 0.05, we say that there is evidence against the null hypothesis in favour of the alternate hypothesis. In another word, we have evidence that the true correlation cannot be zero and hence there exists linear relationship between the two variables.

Example 8.2

```
cor.test(rain_df$PRECIP, rain_df$ALTITUDE)
```

```
    Pearson's product-moment correlation

data:  rain_df$PRECIP and rain_df$ALTITUDE
t = 1.6763, df = 28, p-value = 0.1048
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0653756  0.5972887
sample estimates:
      cor
0.3020067
```

The output showed that the p-value is around 0.10, which is greater than 0.05. This implies that we do not have sufficient evidence against the null hypothesis. In another word, precipitation and altitude are not correlated.

9. 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.

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.

9.1 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.

Example 10.1: Suppose we want to know whether the altitude of the station affect annual precipitation?

- The dependent variable here is annual precipitation, whereas the independent variable is the independent variable.
- To fit this linear regression model in R,

```
model <- lm(PRECIP~ALTITUDE, data=rain_df)
```

Example 9.1: Codes and ouput

```
summary(model)
```

```
Call:
lm(formula = PRECIP ~ ALTITUDE, data = rain_df)

Residuals:
    Min      1Q  Median      3Q     Max
-20.620  -8.479  -2.729   4.555  58.271

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.514799   3.539141   4.666  6.9e-05 ***
ALTITUDE     0.002394   0.001428   1.676    0.105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.13 on 28 degrees of freedom
Multiple R-squared:  0.09121,   Adjusted R-squared:  0.05875
F-statistic:  2.81 on 1 and 28 DF,  p-value: 0.1048
```

This result is the same as that of a correlation test:

```
cor.test(rain_df$PRECIP,rain_df$ALTITUDE)
```

```
    Pearson's product-moment correlation

data:  rain_df$PRECIP and rain_df$ALTITUDE
t = 1.6763, df = 28, p-value = 0.1048
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0653756  0.5972887
sample estimates:
      cor
0.3020067
```

9.1.1 Model diagnostics

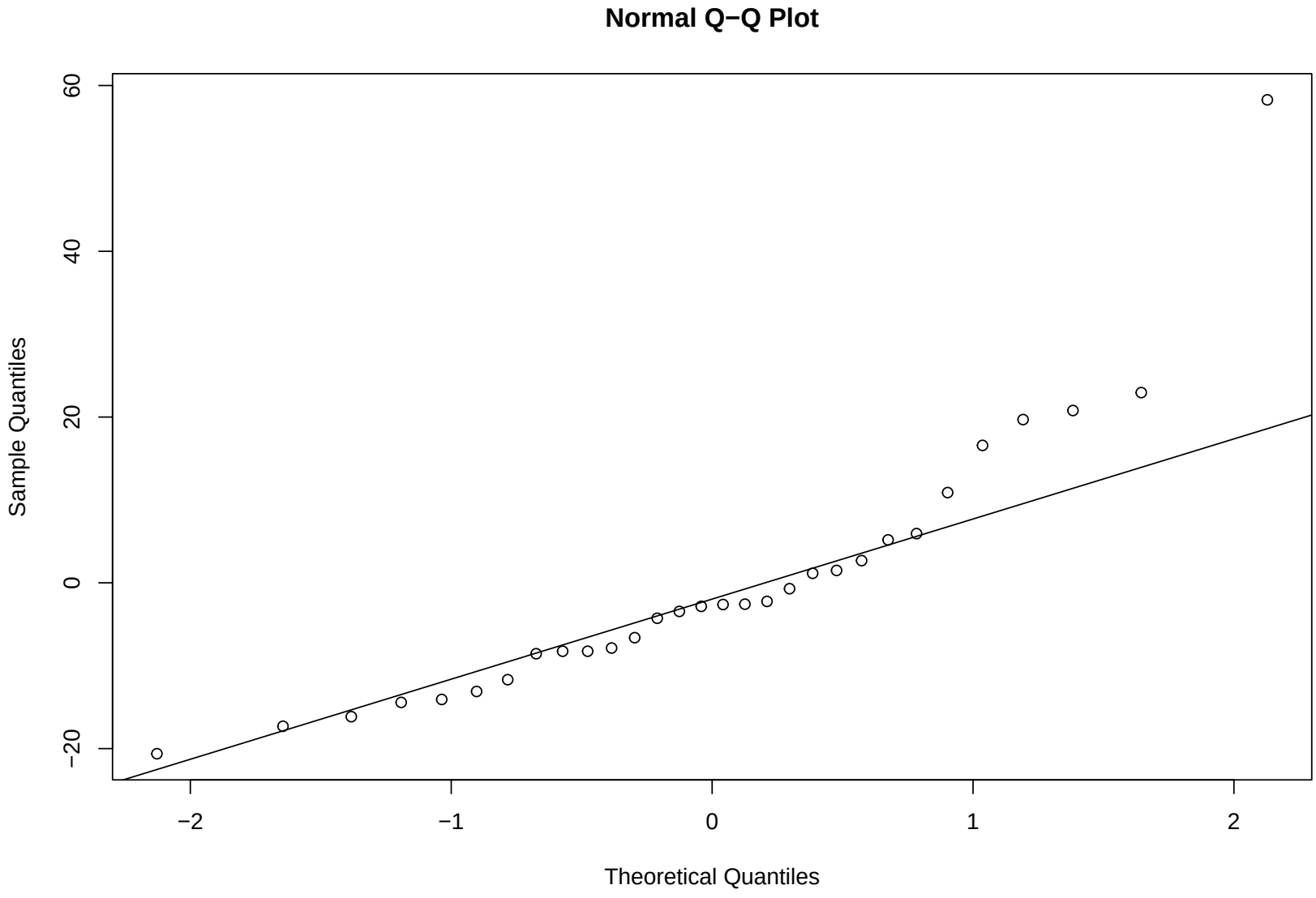
When the assumptions of the linear regression model are satisfied, the model is powerful in terms of inference and interpretation. How do we know whether the assumptions are satisfied?

- Model diagnostics.

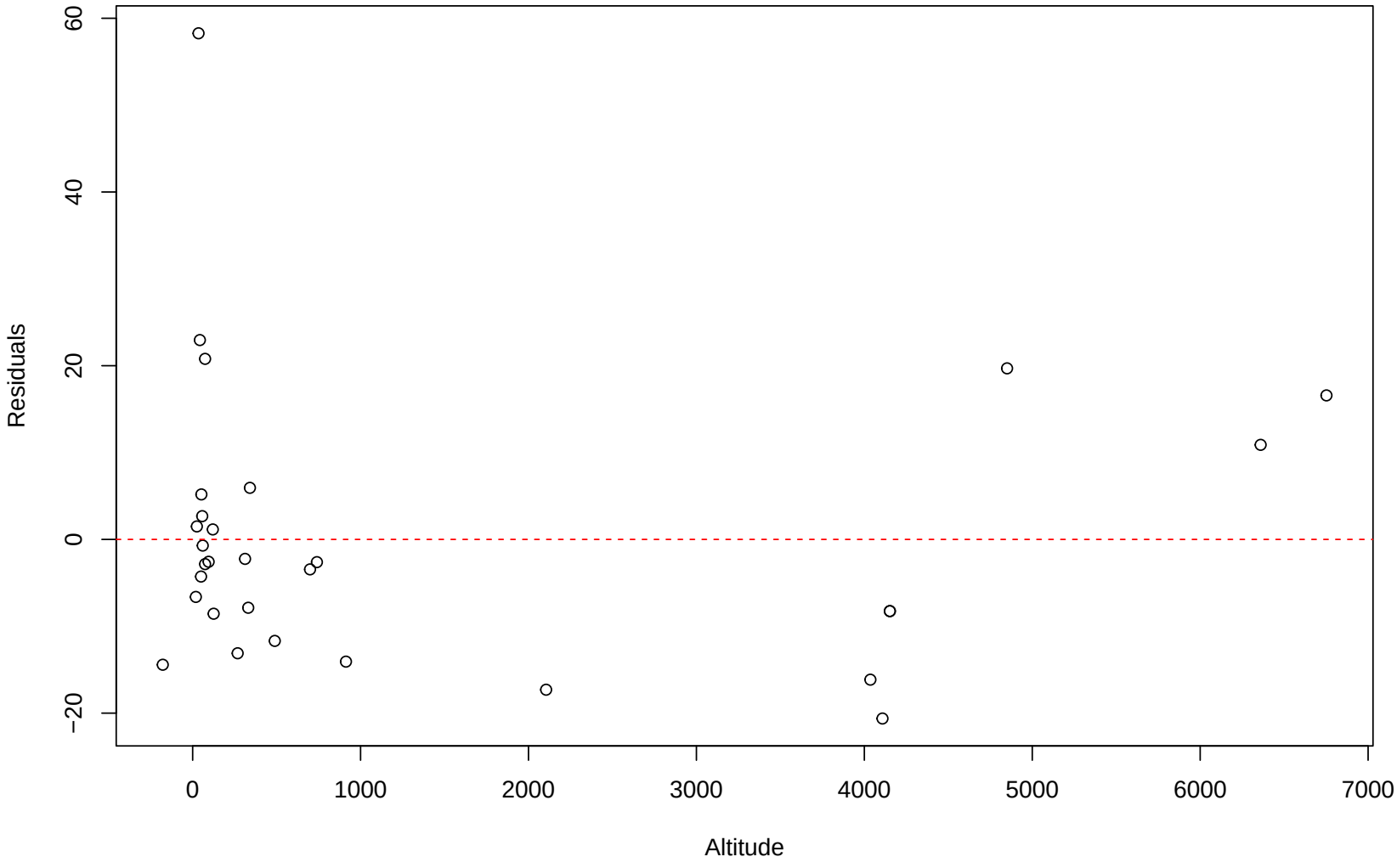
There are a variety of model diagnostics test for different model assumptions. Due to time constraint, we will only cover two simple tools:

- **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 $y = 0$.

QQ-plot for Example 9.1



Residual against explanatory variable for Example 9.1



9.2 Multiple Linear Regression

- In reality, we want to consider the effect of a combination of independent variables.
- The multiple linear regression model is an extension of the simple linear regression model.
- In a multiple linear regression model, the independent variables can have "combined" effects, which can be modeled as "interactions" among variables.
- This is different from multi-variate models. A multi-variate model uses explanatory variable(s) to model multiple response variable simultaneously.

Example 9.2

Suppose we want to know whether ALTITUDE and SHADOW of the station affect annual precipitation?

- The response variable is annual precipitation, PRECIP.
- The independent variables are ALTITUDE and SHADOW.

To provide more context, SHADOW has values 1 and 2 to represent whether the station is located westward or leeward. It is important to set SHADOW as a categorical variable before performing the analysis. Otherwise R will treat all numerical values as numeric.

```
rain_df$SHADOW <- factor(rain_df$SHADOW,
levels=c("1", "2"),
labels=c("Westward", "Leeward"))
```

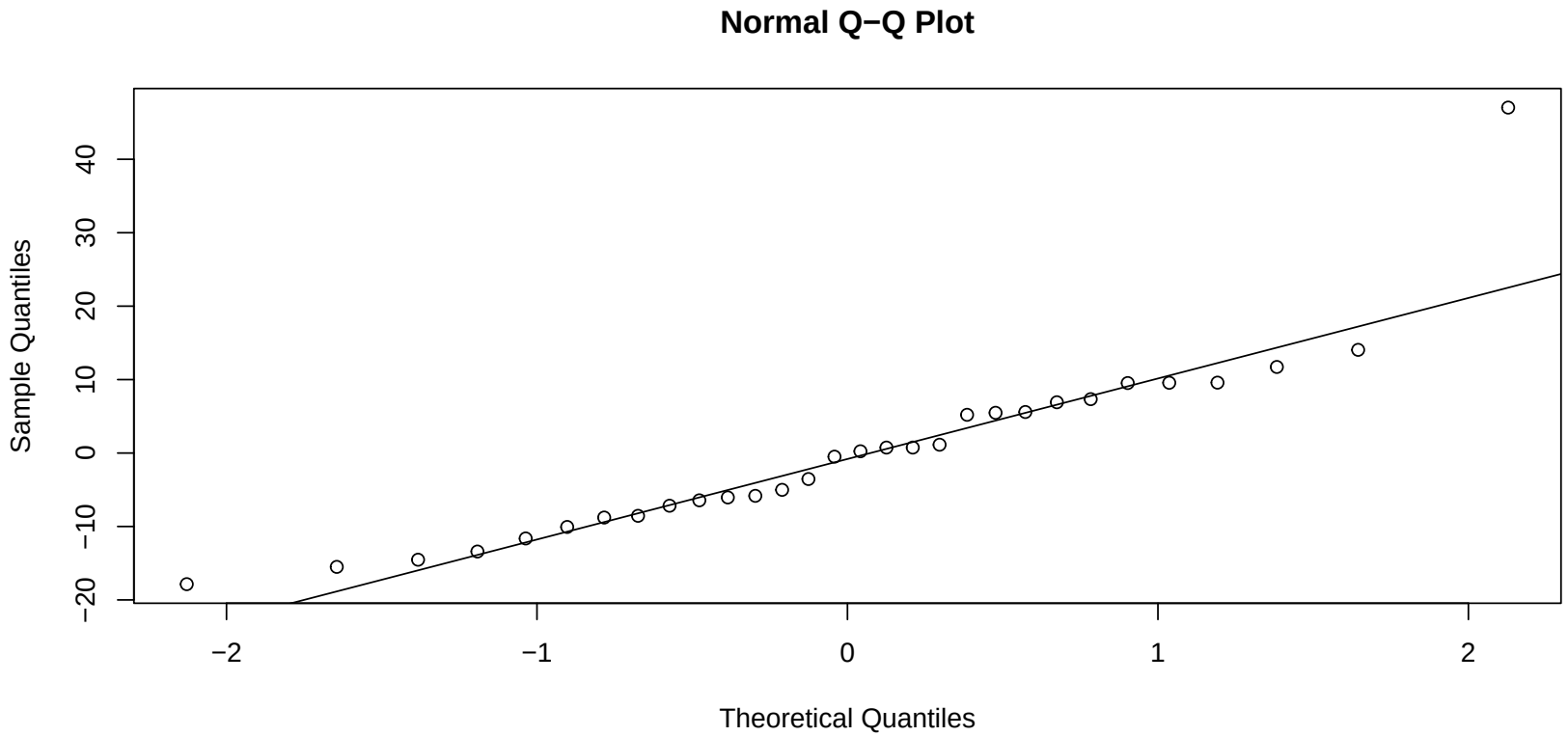
Example 9.2: Analysis

```
model_multiple <- lm(PRECIP~ALTITUDE+SHADOW, data=rain_df)
```

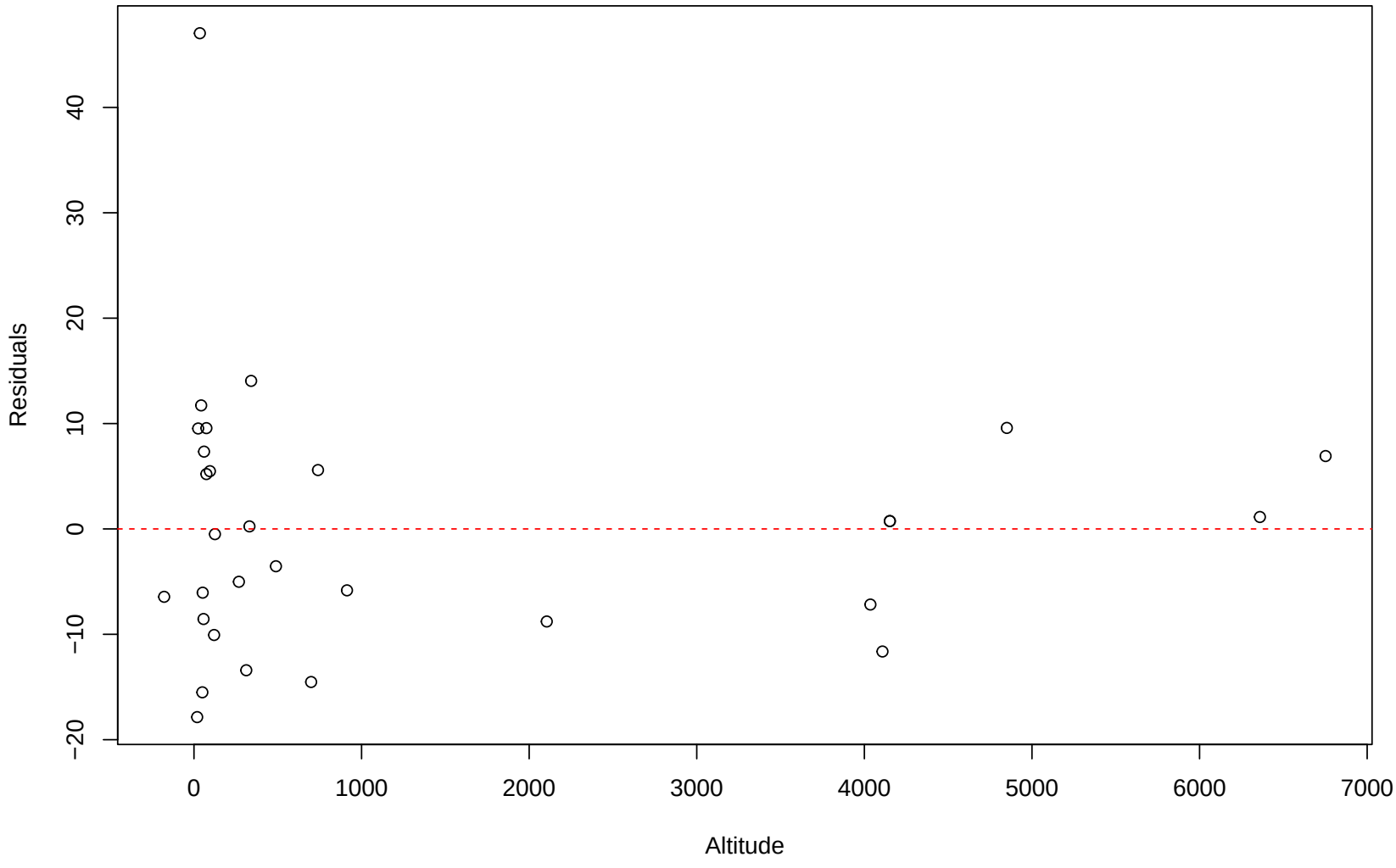
Are you tempted to use the `summary()` function to check the results? Before looking at the output, it is more important to check the fit of the model.

Example 9.2: QQ-plot

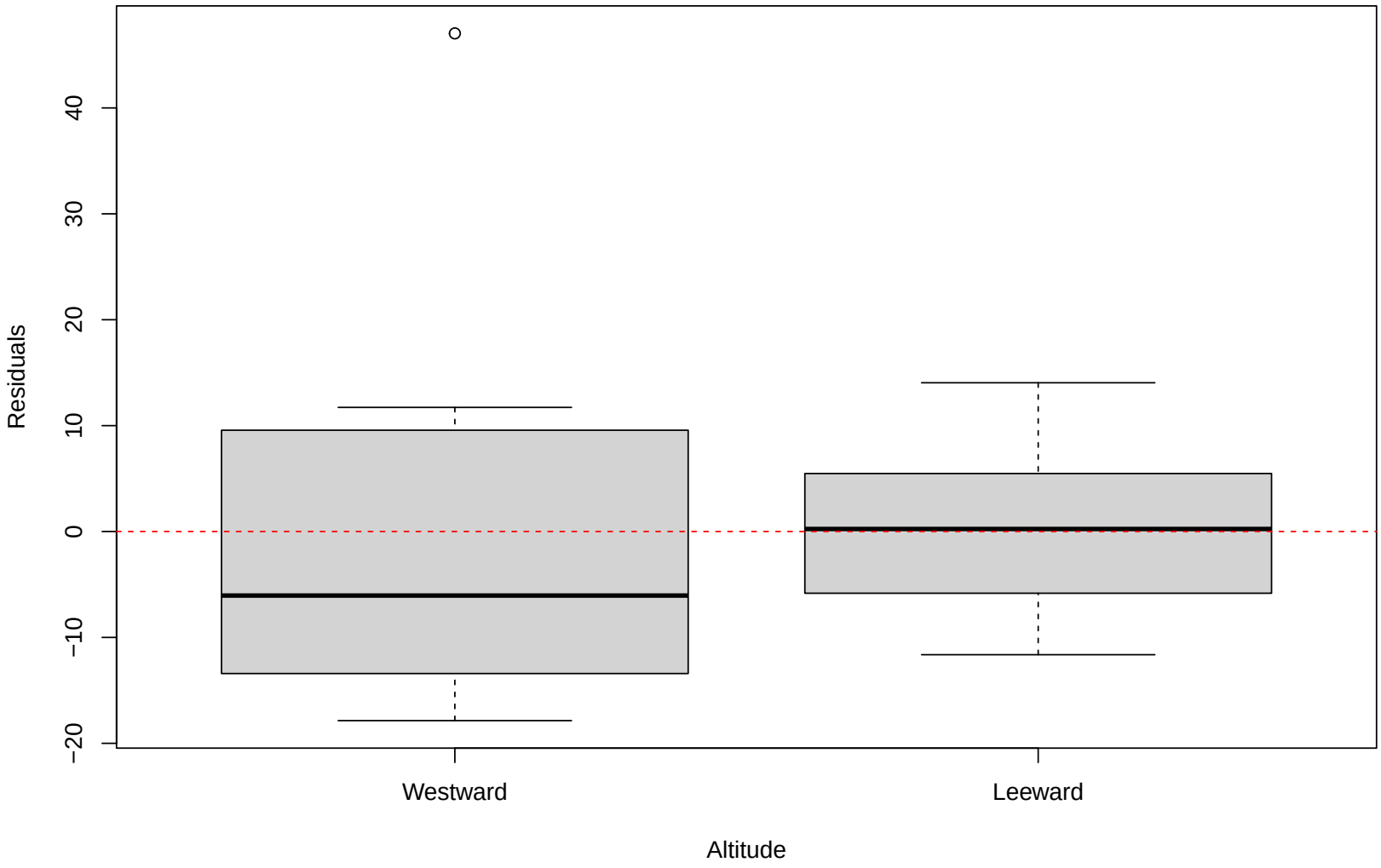
```
qqnorm(model_multiple$residuals)
qqline(model_multiple$residuals)
```



Residual against ALTITUDE for Example 9.1



Residual against SHADOW for Example 9.1



Example 9.2: What did we learn from the plots?

Notice that there is only one point that is far away from the line in the QQ-plot. This is an improvement from the previous linear model. Similar comment can be made about the plot of residual against `SHADOW`. However, `ALTITUDE` does not quite fit with the model.

We can improve the model fit better by considering:

- Adding more independent variables, 2-factor interaction, or higher order terms.
- Removing variables.
- Investigating the point to understand whether it is an outlier or extreme observation.

Due to time constraints, we will leave this model as is and try to interpret the output.

Example 9.2: Output and interpretation

```
summary(model_multiple)
```

```
Call:
lm(formula = PRECIP ~ ALTITUDE + SHADOW, data = rain_df)

Residuals:
    Min      1Q  Median      3Q     Max
-17.857  -8.206  -0.128   6.583  47.039

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    27.755652   3.991116   6.954 1.79e-07 ***
ALTITUDE        0.002161   0.001151   1.876 0.071436 .
SHADOWLeeward -19.270370   4.790149  -4.023 0.000417 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.98 on 27 degrees of freedom
Multiple R-squared:  0.4318,    Adjusted R-squared:  0.3897
F-statistic: 10.26 on 2 and 27 DF,  p-value: 0.0004851
```

Example 9.2: Interpretation

- There is no sufficient evidence that `ALTITUDE` affects precipitation, $\beta = 0.00, t(27) = 1.88, p = .07$.
- There is strong evidence that `SHADOW` has negative association with precipitation, $\beta = -19.27, t(27) = -4.02, p < .001$. Since `SHADOW` is a binary variable, this implies that the `Leeward` precipitation is estimated to be 19.27 less than the baseline, i.e. `Westward`.

9.3 Model fitting procedure

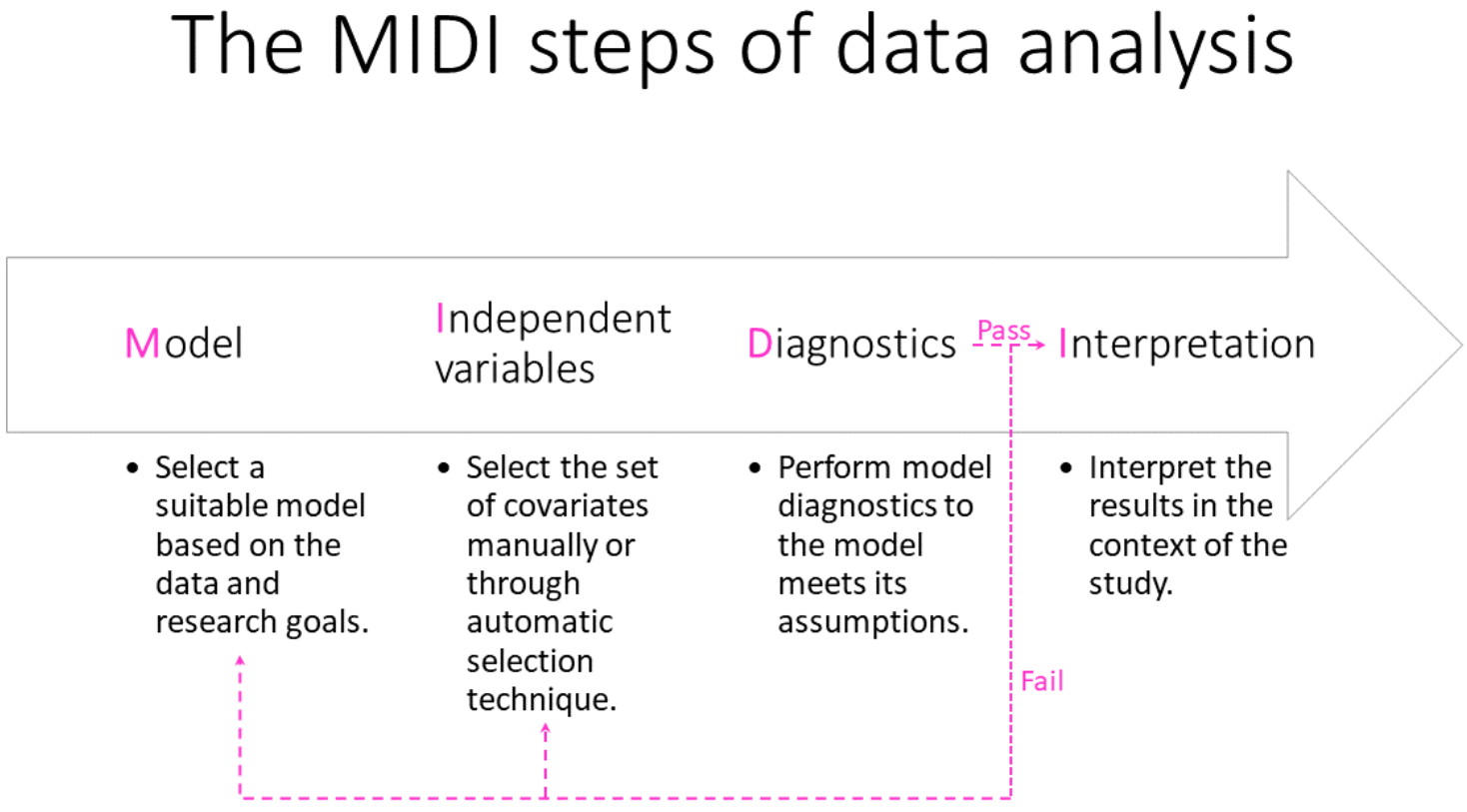


Figure 1: Recommended steps to data analysis

10. 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 SCSR offers free 1-1 consultation to all UWaterloo researchers.

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

11. 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

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

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