Introduction to Business Analytics

Lecture 7: Statistical Methods: Inferences and Regressions

Iegor Vyshnevskyi

Woosong University

April 11/15, 2023

Agenda

1. Intro to Statistical Inference
2. Intro to Confidence Interval
3. Calculation of Confidence Interval in R
4. Basic Concepts of Hypothesis Testing
5. Hypothesis Testing Practical Examples
6. Intro to Regression
7. Assumptions of Linear Regression
8. Regression: practical use
9. In-class Assignment

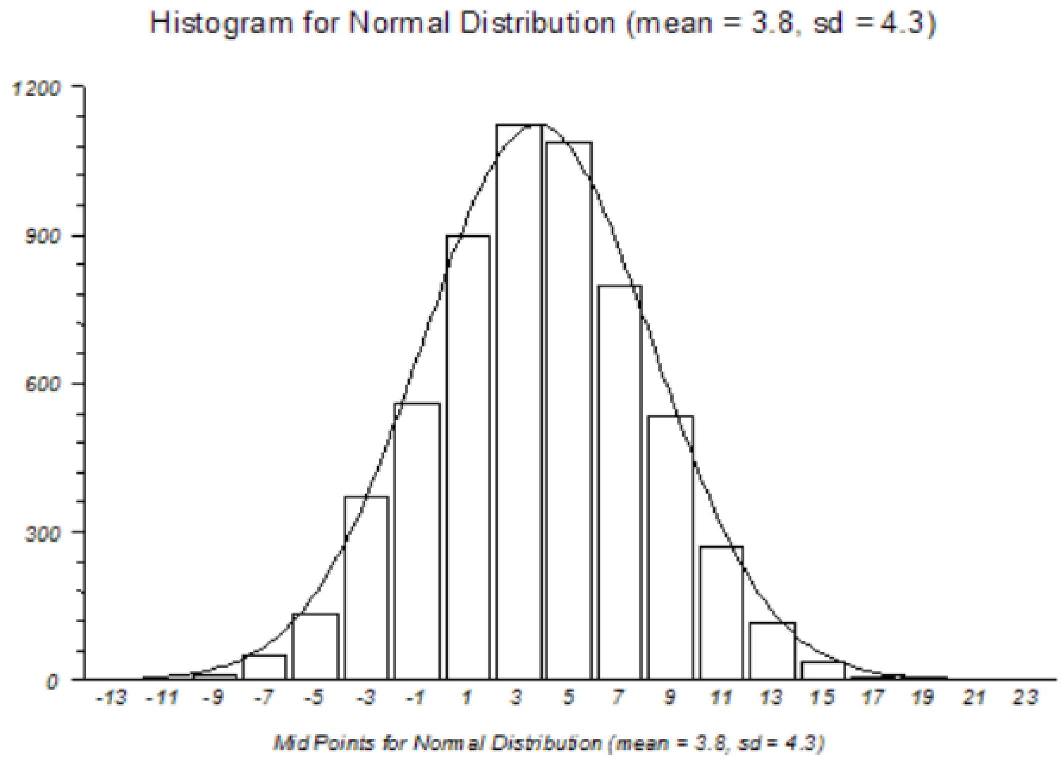
1. Intro to Statistical Inference

Terminology

- A ***variable*** is what we measure and want to study in our project. It could be employee salaries or the transaction value of customers, for example
- A ***population*** is a set of all units we want to draw conclusions about. For example: All the employees in an organization.
- And a ***sample*** is a subset of employees (in other words a specific group of employees) in the organization.

Terminology (cont.)

- A statistical distribution gives us an idea about how these values are distributed in a population. The most common distribution is a ***normal distribution***.



- A ***factor*** defines sub groups in a study such as the gender or location of employees.
- ***Descriptive Statistics*** typically include the mean, median, and standard deviation of a variable under study.

Statistical Inference

the process of drawing conclusions about unknown population properties, using a sample drawn from the population.

Unknown population properties can be, for example, mean, proportion or variance. These are also called ***parameters***.

Type of Statistical Inference

- ***statistical estimation*** is concerned with best estimating a value or range of values for a particular population parameter, and
- ***hypothesis testing*** is concerned with deciding whether the study data are consistent at some level of agreement with a particular population parameter.

Resource: https://www.engati.com/glossary/statistical-inference

Type of Statistical Inference (cont.)

Statistical Inference

Estimation

Point estimation

- Summarize the sample by a single value as an estimate of the true population parameter.
- Ex. Average salary of junior data scientists is. 55,000 euros.

Interval estimation

- A range of values within which, we believe,
  the true population parameter lies with high probability.
- Ex. Average salary of junior data scientists is in the range of (52,000,55,000) With 95% confidence level.

Testing of Hypothesis

- To decide whether a statement regarding population parameter is true or false, based on sample data.
- Ex. Claim: Average salary of junior data scientists is greater than.50,000 euros annually.

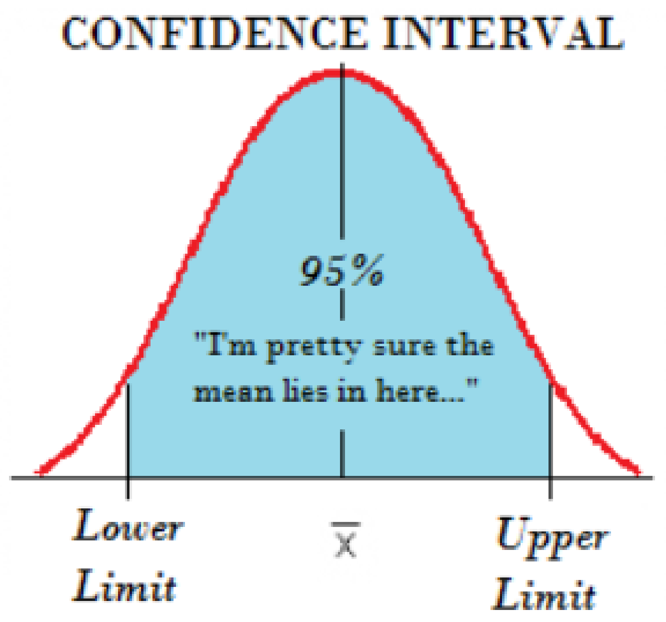
Resource: https://digitaschools.com/what-is-statistical-inference/

2. Intro to Confidence Interval

Confidence Interval

- the mean of your estimate plus and minus the variation in that estimate. This is the range of values you expect your estimate to fall between if you redo your test, within a certain level of confidence.

- ***Confidence***, in statistics, is another way to describe probability. For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.



Resource: https://www.statisticshowto.com/confidence-limits-definition/

Confidence Interval Formula

Confidence interval =

Mean of sample ± Test Statistic * Standard Error

Test Statistic is a number calculated from a statistical test of a hypothesis. It shows how closely your observed data match the distribution expected under the null hypothesis of that statistical test.

Standard error = standard deviation / squared number of observations

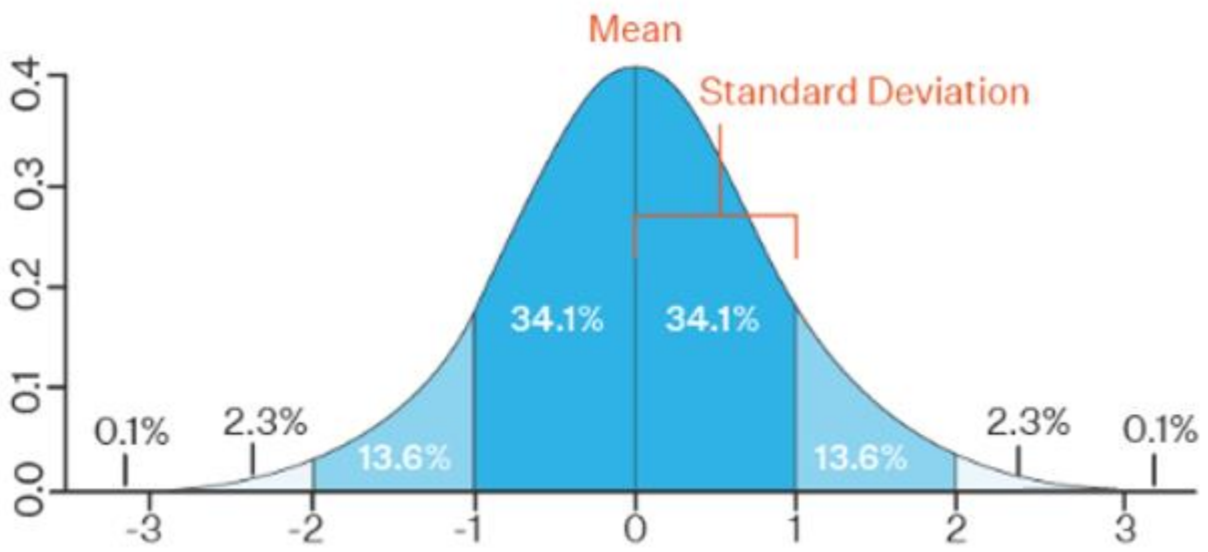
$$\text{Standard error} = \frac{\sigma_x}{\sqrt{N}}$$

Type of Most Common Test Statistics

Test statistic	Null and alternative hypotheses	Statistical tests that use it
t value	**Null:** The means of two groups are equal **Alternative:** The means of two groups are not equal	• *T* test • Regression tests
z value	**Null:** The means of two groups are equal **Alternative:**The means of two groups are not equal	• *Z* test
F value	**Null:** The variation among two or more groups is greater than or equal to the variation between the groups **Alternative:** The variation among two or more groups is smaller than the variation between the groups	• ANOVA • ANCOVA • MANOVA
X^2-value	**Null:** Two samples are independent **Alternative:** Two samples are not independent (i.e., they are correlated)	• Chi-squared test • Non-parametric correlation tests

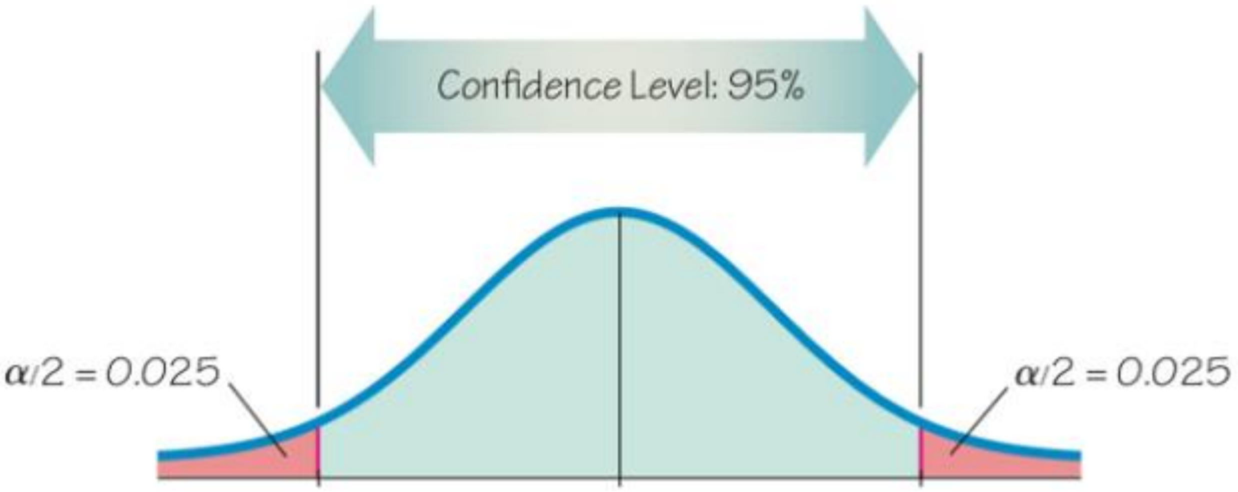
Resource: https://www.scribbr.com/statistics/test-statistic/

Standard deviation



- Around 68% of scores are within 1 standard deviation of the mean,
- Around 95% of scores are within 2 standard deviations of the mean,
- Around 99.7% of scores are within 3 standard deviations of the mean.

Standard deviation



Confidence Level	Alpha	Alpha/2
90%	10%	5.0%
95%	5%	2.5%
98%	2%	1.0%
99%	1%	0.5%

Resource: https://math.stackexchange.com/questions/2835809/one-tailed-confidence-interval-1-2-alpha-rationale
https://www.statisticshowto.com/probability-and-statistics/confidence-interval/

Degree of Freedom

- the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample.

$$D_f = N - 1$$

where:

D_f = degrees of freedom

N = sample size

3. Calculation of Confidence Interval in R

```r
library(tidyverse)
data(iris)
head(iris)
```

- The iris dataset is a famous multivariate dataset in R that contains measurements for different parts of flowers belonging to three different species of iris: setosa, versicolor, and virginica.
- The iris dataset contains 150 observations, with 50 observations for each of the three iris species. For each observation, the following four measurements are recorded: Sepal length (in cm), Sepal width (in cm), Petal length (in cm), Petal width (in cm).

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

Let’s calculate the 95% confidence interval for population mean of Sepal.Length.
Recall the formula of confidence interval.

Confidence interval =

Mean of sample ± Test Statistic * Standard Error

So, to calculate the confidence interval first we need to calculate:

- Mean
- Test statistic (t-value in our case)
- Standard error

Calculation of the sample mean

```r
# Calculate the mean of the sample data
mean_value <- mean(iris$Sepal.Length)
mean_value
```

```
> mean_value
[1] 5.843333
```

Calculation of the test statistics (t-value)

```
# Compute the size of the sample
n <- length(iris$Sepal.Length)
n
```

Result:

```
> n
[1] 150
```

```
#assign the alpha
alpha <- 0.05

# Compute the degree of freedom
degrees_of_freedom <- n - 1
degrees_of_freedom
```

Result:

```
> degrees_of_freedom
[1] 149
```

Calculation of the test statistics (t-value) (cont.)

```
# use the function qt() to calculate the t-score for
#a given level of significance (alpha) and degrees of freedom.
t_score <- qt(p = alpha/2,
              df = degrees_of_freedom,
              #to calculate the t-value corresponding to
              #the upper tail of the t-distribution
              lower.tail=F)
t_score
```

Result:

```
> t_score
[1] 1.976013
```

For lower tail the t-value will be -1.97.

Calculation of standard error

```
# Find the standard deviation
standard_deviation <- sd(iris$Sepal.Length)
standard_deviation
```

Result:

```
> standard_deviation
[1] 0.8280661
```

```
# Find the standard error
standard_error <- standard_deviation / sqrt(n)
standard_error
```

Result:

```
> standard_error
[1] 0.06761132
```

Calculation the confidence interval

```
# Calculating lower bound and upper bound
lower_bound <- mean_value - t_score * standard_error
upper_bound <- mean_value + t_score * standard_error

# Print the confidence interval
print(c(lower_bound,upper_bound))
```

Result:

```
> print(c(lower_bound,upper_bound))
[1] 5.709732 5.976934
```

4. Basic Concepts of Hypothesis Testing

*What is **Hypothesis Testing***

a type of statistical analysis in which you put your assumptions about a population parameter to the test.

First, we need to state the ***null*** and ***alternative*** hypothesis.

The ***Alternative Hypothesis*** is the Hypothesis which we are maintaining and would like to prove.

Example question: A company claims that their new product is more effective than the current market leader.

> ***Null hypothesis:*** The new product is **not** more effective than the current market leader.
>
> ***Alternative hypothesis***: The new product is more effective than the current market leader.

Example question: A researcher wants to investigate if there is a difference in the mean weight of two different breeds of dogs.

> ***Null hypothesis:*** There is **no** significant difference in the mean weight of the two breeds of dogs.
>
> ***Alternative hypothesis:*** There is a significant difference in the mean weight of the two breeds of dogs.

Null vs. Alternative Hypothesis

Null Hypothesis

H_0

A statement about a population parameter.

We test the likelihood of this statement being true in order to decide whether to accept or reject our alternative hypothesis.

Can include =, ≤, or ≥ sign.

Alternative Hypothesis

H_a

A statement that directly contradicts the null hypothesis.

We determine whether or not to accept or reject this statement based on the likelihood of the null (opposite) hypothesis being true.

Can include a ≠, >, or < sign.

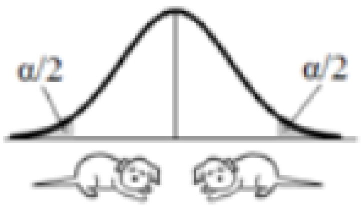
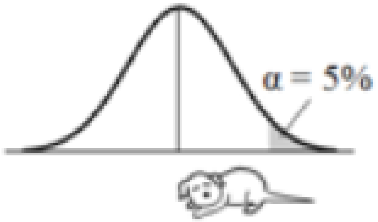
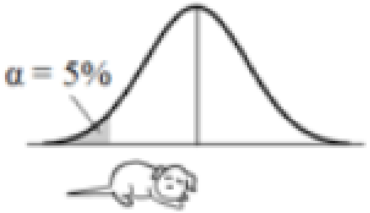
ThoughtCo.

Type of hypothesis tests (cheat sheet)

Type Of Test	Purpose	Example
Z Test	Test if the average of a single population is equal to a target value	Do babies born at this hospital weigh more than the city average
1 Sample T-Test	Test if the average of a single population is equal to a target value	Is the average height of male college students greater than 6.0 feet?
Paired T-Test	Test if the average of the differences between paired or dependent samples is equal to a target value	Weigh a set of people. Put them on a diet plan. Weigh them after. Is the average weight loss significant enough to conclude the diet works?
2 Sample T-Test Equal Variance	Test if the difference between the averages of two independent populations is equal to a target value	Do cats eat more of type A food than type B food
2 Sample T-Test Unequal Variance	Test if the difference between the averages of two independent populations is equal to a target value	Is the average speed of cyclists during rush hour greater than the average speed of drivers

Resource: https://studylib.net/doc/25486595/hypothesis-testing-cheat-sheet-1

One and Two-tailed tests

Comparison Operator		Tails of the Test	
H_A	H_0		
$\neq$	$=$	2-tailed	
$>$	$\leq$	1- tailed, right-tailed	
$<$	$\geq$	1-tailed, left-tailed	

Example:

To test whether the Mean lifetime of the lightbulbs we manufacture is more than 1,300 hours.

For two-tailed test:

H0: Mean = 1300

Ha: Mean ≠ 1300

For right tailed test:

H0: Mean ≤ 1300

Ha: Mean > 1300

For left tailed test:

H0: Mean ≥ 1300

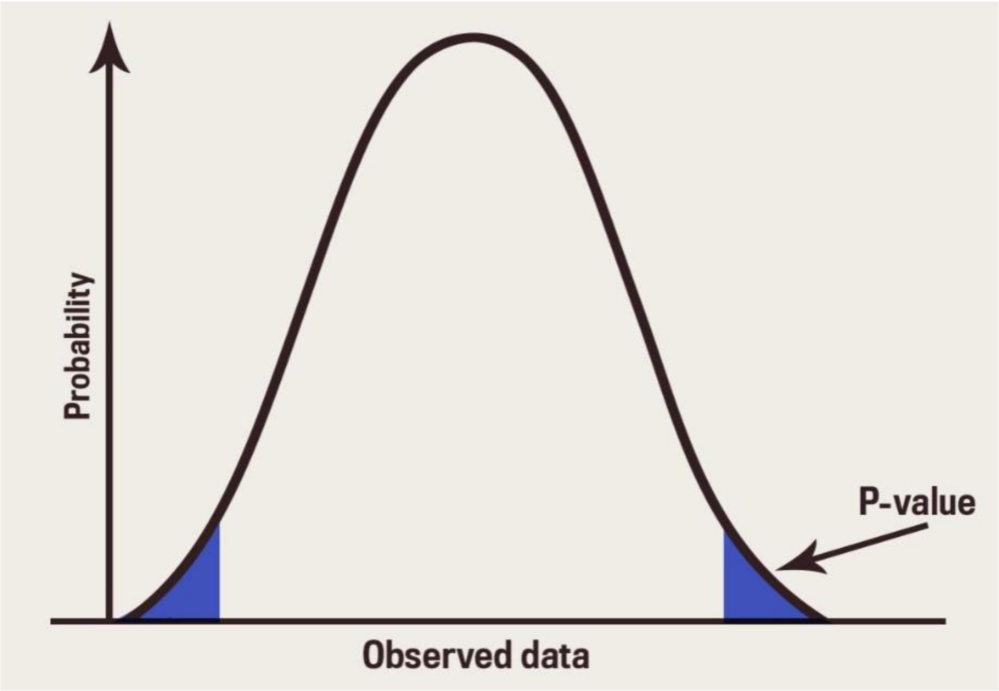
Ha: Mean < 1300

P-value

The p-value is a probability.

When the p-value is very small, it means it is very unlikely (small probability) that the observed spatial pattern is the result of random processes, so you can reject the null hypothesis.

An easier way to remember the decision of a hypothesis test is by using the phrase ***"when p is low, the null must go."***

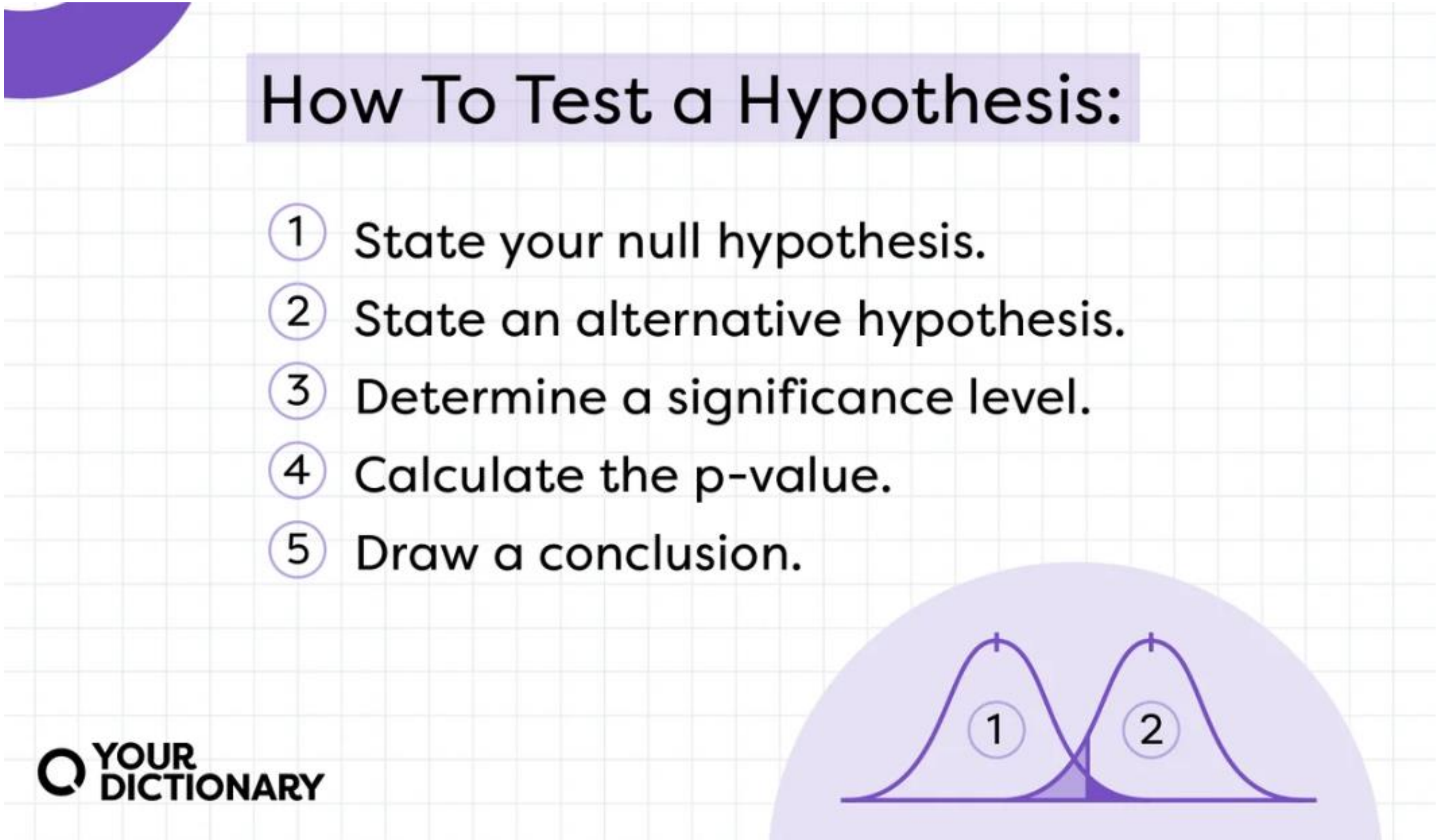


P-value	Decision
Less than 0.05*	**Reject Null** (H_0) Hypothesis Statistical difference between groups
Greater than 0.05*	**Fail to Reject** Null (H_0) Hypothesis No statistical difference between groups, or not enough evidence (data) to find a difference

* Assuming $\alpha = 0.05$

Resource: https://www.leansixsigmadefinition.com/glossary/p-value/

To sum up


How To Test a Hypothesis:
1 State your null hypothesis.
2 State an alternative hypothesis.
3 Determine a significance level.
4 Calculate the p-value.
5 Draw a conclusion.
1
2
YOUR DICTIONARY

5. Hypothesis Testing Practical Examples

State the null and alternative hypothesis

Going back to our iris example, consider the situation where we wish to determine whether the mean Sepal Length is 6 or not at the 0.05 level of significance.

Null hypothesis: The mean Sepal Length of the iris population is equal to 6.

Alternative hypothesis: The mean Sepal Length of the iris population is not equal to 6.

Two-tailed test

```r
library(stats)

#test whether mean Sepal.Length is equal to 6
t.test(x = iris$Sepal.Length,
       alternative = "two.sided",
       mu = 6)
```

Result:

```
        One Sample t-test

data:  iris$Sepal.Length
t = -2.3172, df = 149, p-value = 0.02186
alternative hypothesis: true mean is not equal to 6
95 percent confidence interval:
 5.709732 5.976934
sample estimates:
mean of x
 5.843333
```

Conclusion:

p-value < 0.05 level of significance,

which suggests that the null hypothesis can be rejected in favor of the alternative hypothesis.

The mean Sepal.Lenght is NOT equal to 6 at 0.05 level of significance.

One-tailed test (right-tailed test)

```r
#test whether mean Sepal.Length >= 6 (right-tailed test)
t.test(x = iris$Sepal.Length, alternative = "less",
       mu = 6)
```

Result:

```
        One Sample t-test

data:  iris$Sepal.Length
t = -2.3172, df = 149, p-value = 0.01093
alternative hypothesis: true mean is less than 6
95 percent confidence interval:
    -Inf 5.95524
sample estimates:
mean of x
 5.843333
```

Conclusion:

p-value < 0.05 level of significance,

which suggests that the null hypothesis can be rejected in favor of the alternative hypothesis.

The mean Sepal.Lenght is less than 6 at 0.05 level of significance.

One-tailed test (left-tailed test)

```r
#test whether mean Sepal.Length <= 6 (left-tailed test)
t.test(x = iris$Sepal.Length, alternative = "greater",
       mu = 6)
```

Result:

```
        One Sample t-test

data:  iris$Sepal.Length
t = -2.3172, df = 149, p-value = 0.9891
alternative hypothesis: true mean is greater than 6
95 percent confidence interval:
 5.731427      Inf
sample estimates:
mean of x
 5.843333
```

Conclusion:

p-value > 0.05 level of significance,

which suggests that the null hypothesis can NOT be rejected in favor of the alternative hypothesis.

The mean Sepal.Lenght is NOT greater than 6 at 0.05 level of significance.

Two-sample test

```r
#test whether difference in means of Sepal.Length and Petal.Length is equal to 0
t.test(x = iris$Sepal.Length,
       y = iris$Petal.Length,
         alternative = "two.sided",
       mu = 0)
```

Result:

```
        Welch Two Sample t-test

data:  iris$Sepal.Length and iris$Petal.Length
t = 13, df = 212, p-value <0.0000000000000002
alternative hypothesis: true difference in means is
 not equal to 0
95 percent confidence interval:
 1.772 2.399
sample estimates:
mean of x mean of y
    5.843     3.758
```

Conclusion:

p-value < 0.05 level of significance,

which suggests that the null hypothesis can be rejected in favor of the alternative hypothesis.

difference in means of Sepal.Length and Petal.Length is not equal to 0 at 0.05 level of significance.

6. Intro to Regression

Regression

is a statistical method used to study the relationship between a dependent variable (usually denoted as Y) and one or more independent variables (usually denoted as X).

- It is commonly used for predictive modeling, to estimate the value of the dependent variable based on the values of one or more independent variables.
- The goal of regression is to find the best-fitting line or curve that can describe the relationship between the variables, which can then be used to make predictions or to understand how changes in the independent variable(s) affect the dependent variable.

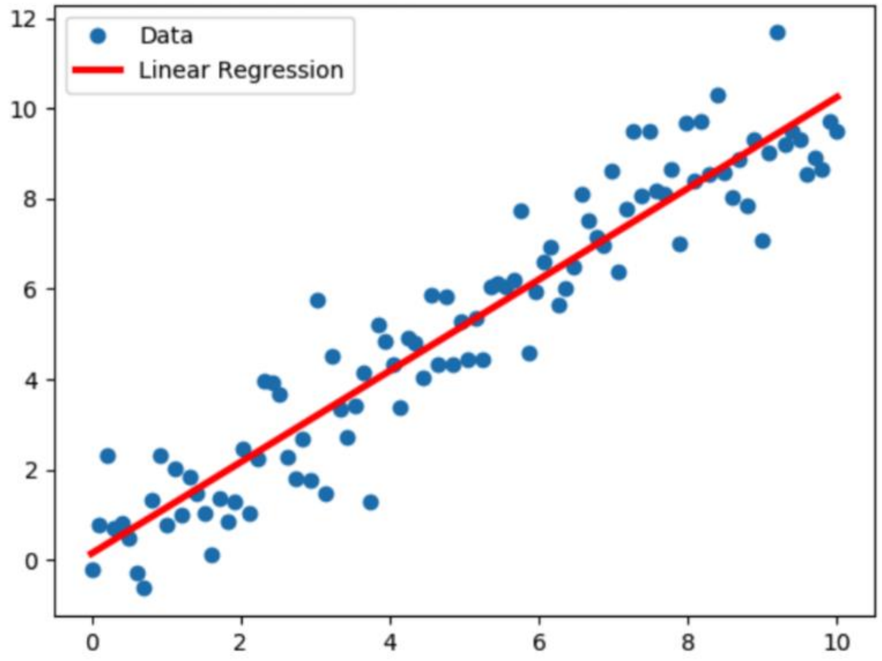
Type of regressions

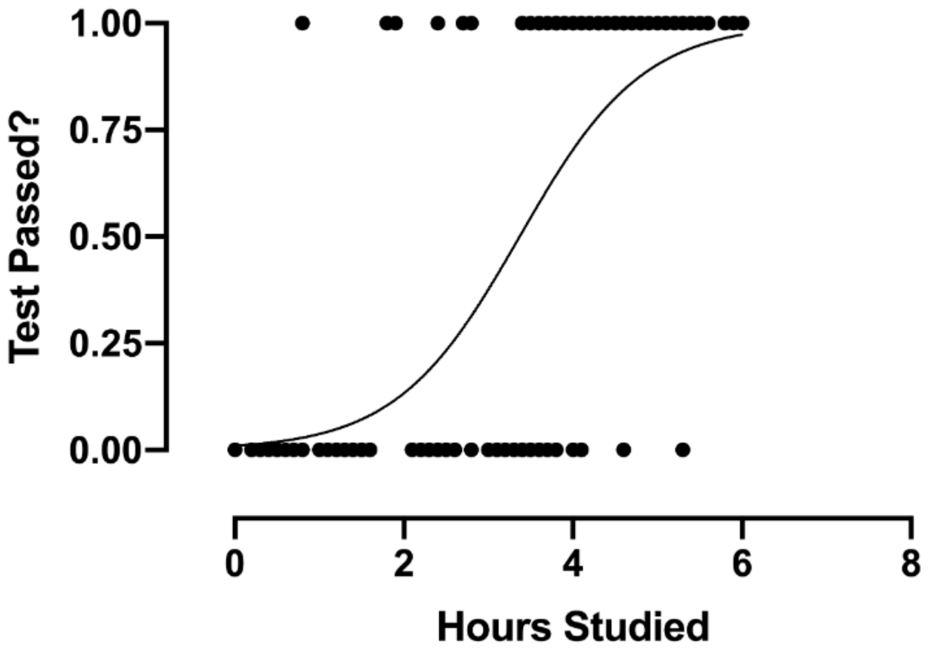
- Linear Regression
- Logistic Regression
- Polynomial Regression
- Ridge Regression
- Lasso Regression
- Quantile Regression
- Bayesian Linear Regression
- Principal Components Regression
- Partial Least Squares Regression
- Elastic Net Regression

Type of regressions (cont.)

Linear Regression: A method that models the relationship between a dependent variable and one or more independent variables as a straight line.

Logistic Regression: A method used to model the probability of a binary outcome (i.e., yes or no, true or false) based on one or more predictor variables.

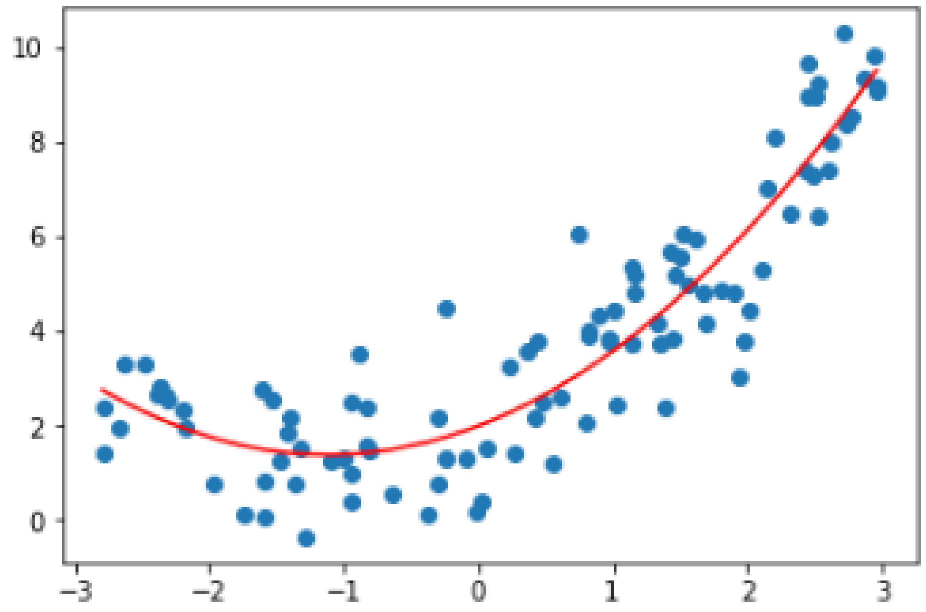




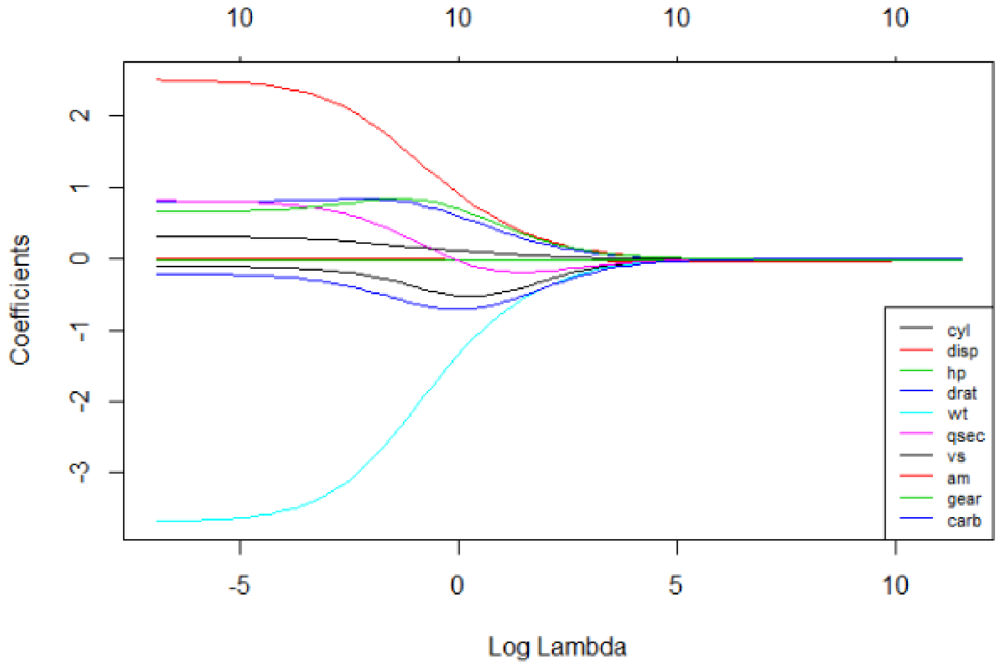
Resource: https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/

Type of regressions (cont.)

Polynomial Regression: A method that models the relationship between a dependent variable and one or more independent variables using a polynomial function.



Ridge Regression: A method used to avoid overfitting in linear regression by adding a penalty term to the regression equation.

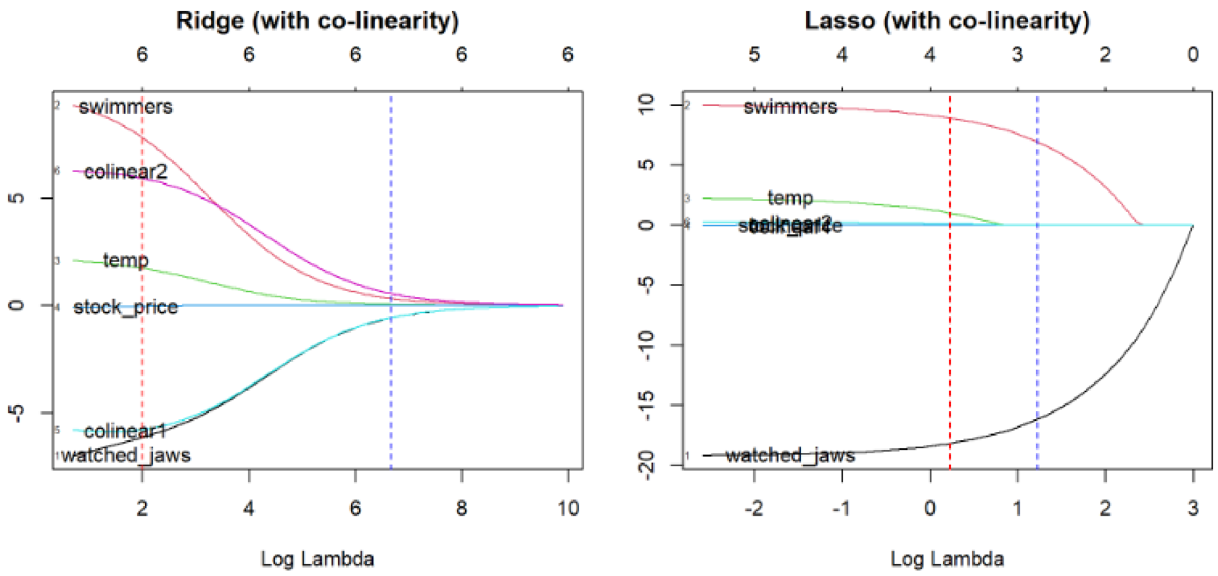


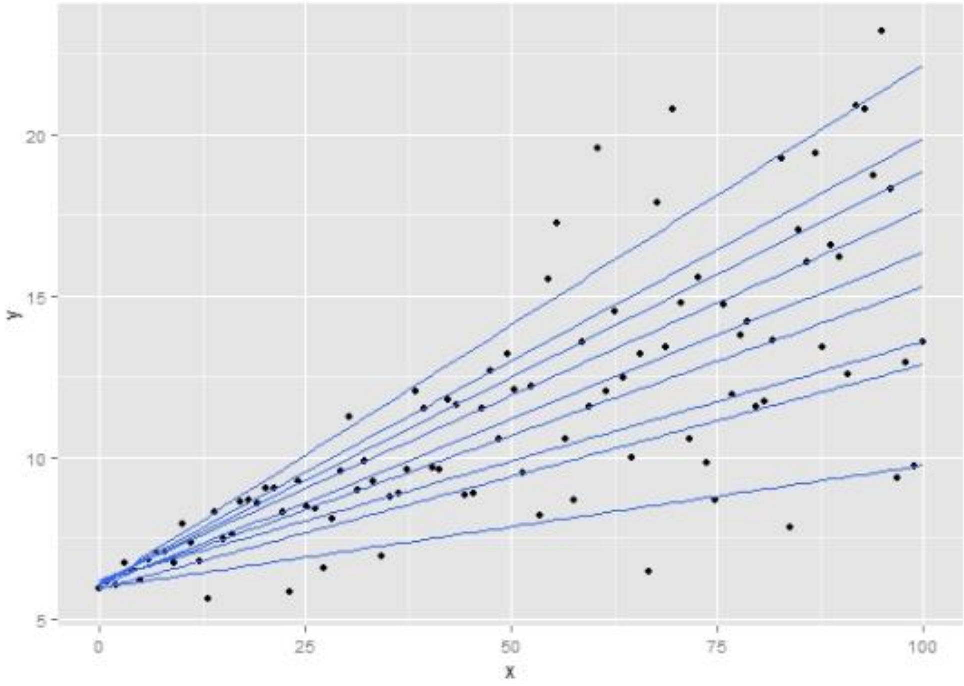
Resource: https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/

Type of regressions (cont.)

Lasso Regression: A method used to select important predictor variables and avoid overfitting in linear regression by shrinking the coefficients of less important variables to zero.

Quantile Regression: A method that estimates the relationship between a dependent variable and one or more independent variables at different quantiles of the dependent variable.

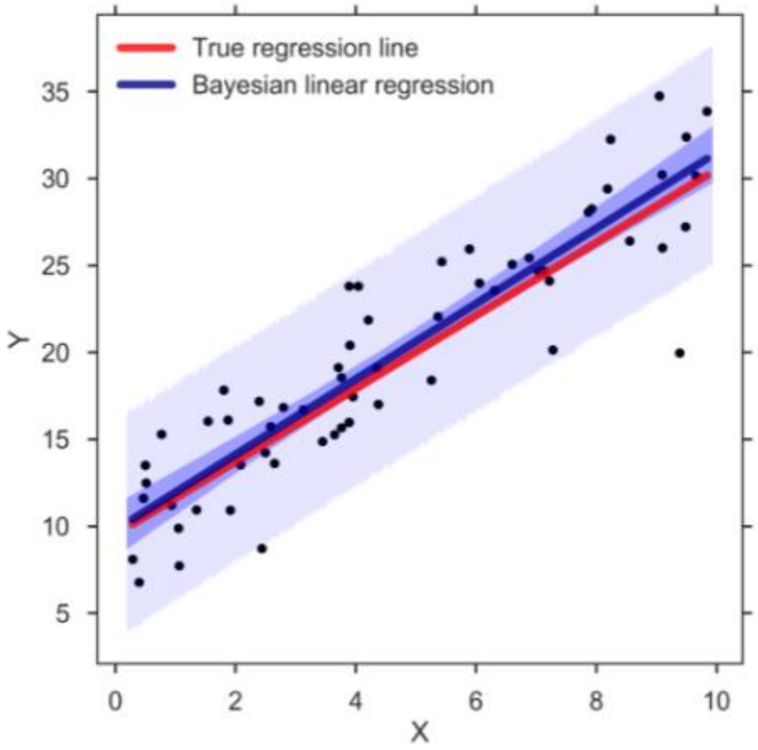




Resource: https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/

Type of regressions (cont.)

Bayesian Linear Regression: A method that uses Bayesian inference to estimate the parameters of a linear regression model.



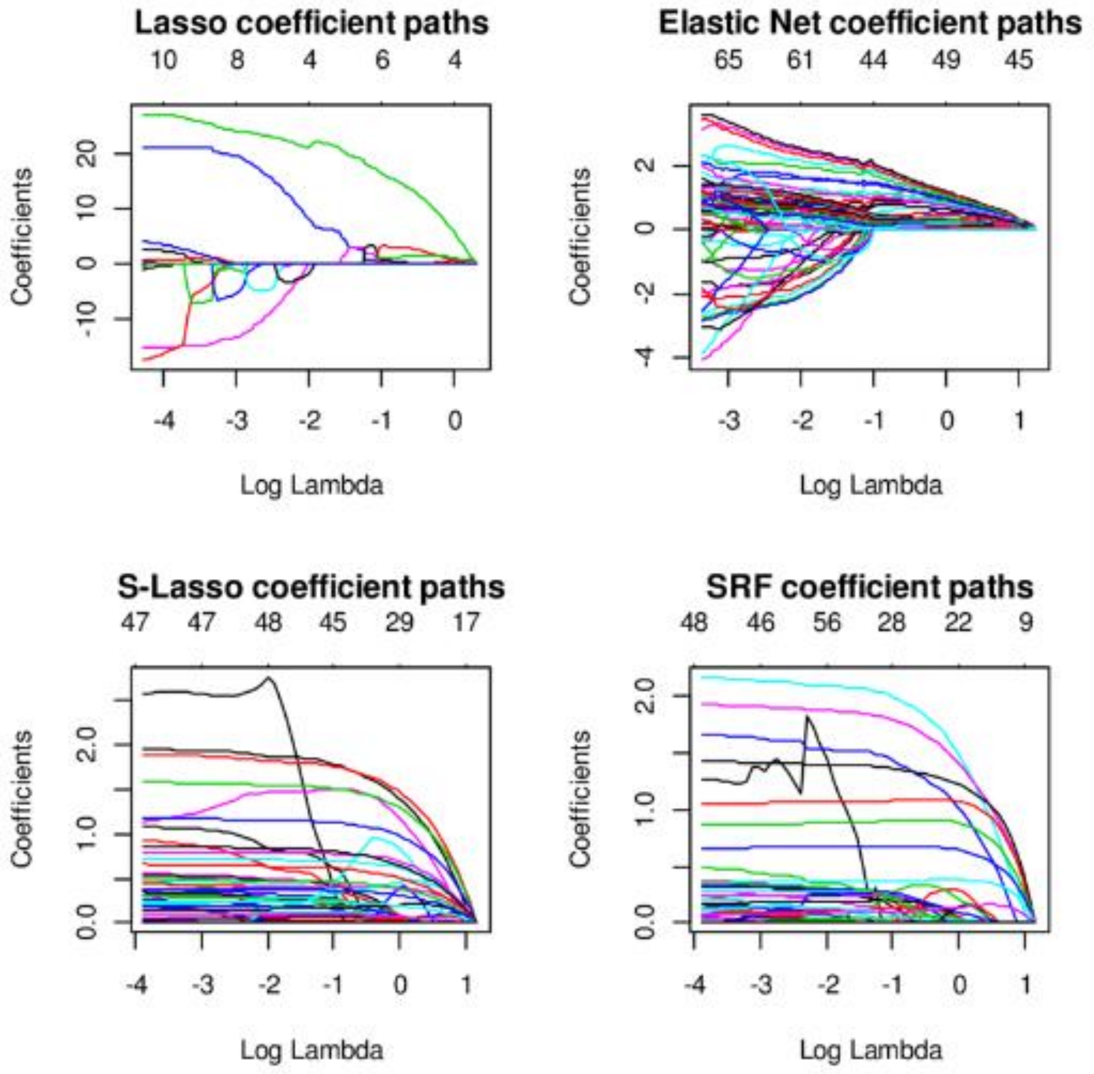
Principal Components Regression: A method that uses principal component analysis to reduce the dimensionality of the predictor variables before performing linear regression.

Partial Least Squares Regression: A method that uses partial least squares regression to reduce the dimensionality of the predictor variables before performing linear regression.

Resource: https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/

Type of regressions (cont.)

Elastic Net Regression: A method that combines the penalty terms of ridge regression and lasso regression to overcome their limitations and select important predictor variables while avoiding overfitting.



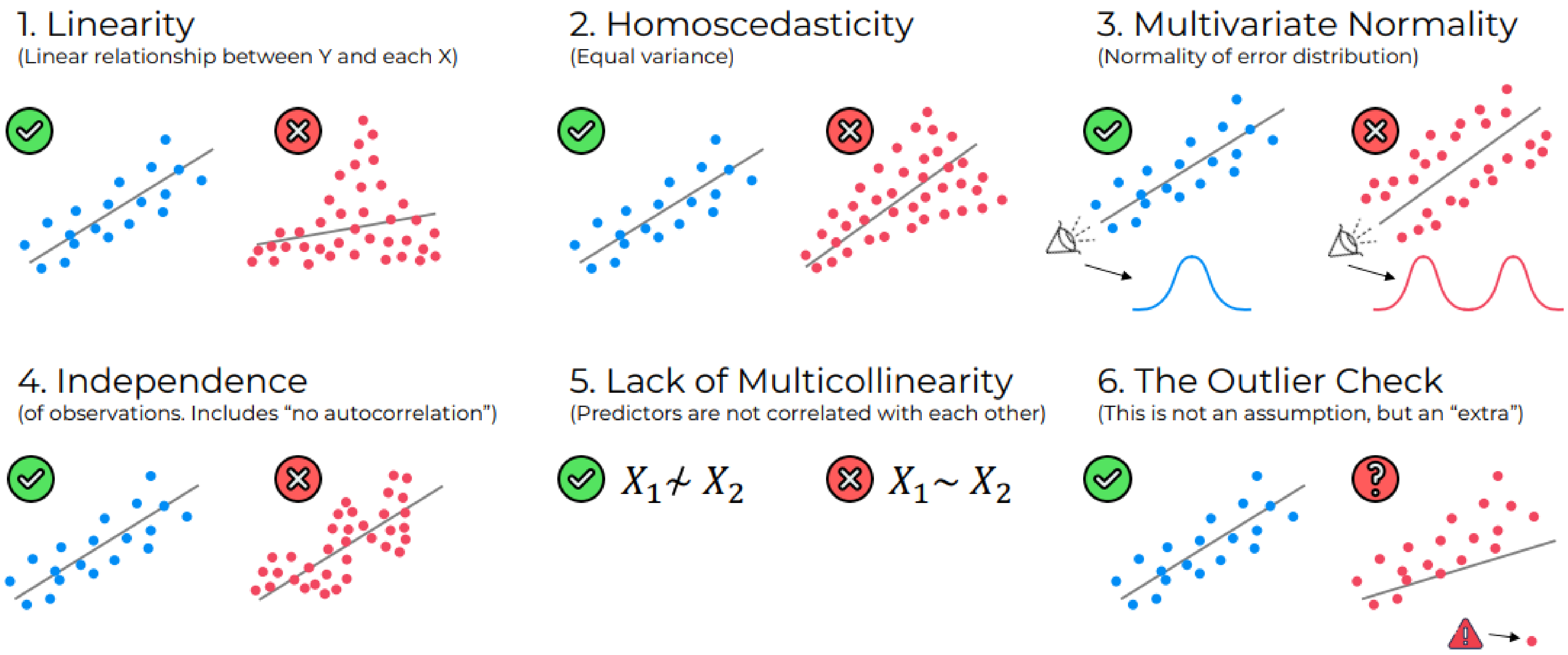
Resource: https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/

7. Assumptions of Linear Regression

Linear regression is a widely used statistical technique for modeling the relationship between a dependent variable and one or more independent variables.

However, the accuracy of the regression model depends on several assumptions that need to be satisfied for the model to be valid.

Key Assumptions of Linear Regression



If these assumptions are not met, the regression model may produce biased or inconsistent estimates, and the results may be invalid.

Therefore, it is important to check for these assumptions before using linear regression for modeling the data.

Resource: https://www.superdatascience.com/blogs/assumptions-of-linear-regression

8. Regression: Practical Use

Let's open the file "Taiwan_data" and quickly look at it.

```
> head(taiwan_data)
  dist_to_mrt_m n_convenience house_age_years price_twd_msq
1         84.88            10        30 to 45        11.467
2        306.59             9        15 to 30        12.769
3        561.98             5         0 to 15        14.312
4        561.98             5         0 to 15        16.581
5        390.57             5         0 to 15        13.041
6       2175.03             3         0 to 15         9.713
```

The table Taiwan_data contains information related to real estate properties in Taiwan.

Here is a brief description of each column:

- dist_to_mrt_m: The distance of the property to the nearest Mass Rapid Transit (MRT) station in meters.
- n_convenience: The number of convenience stores located near the property.
- house_age_years: The age of the property in years.
- price_twd_msq: The price of the property per square meter in New Taiwan Dollars (TWD).

Simple linear regression

```
#Run a linear regression of price_twd_msq vs. n_convenience
mdl_price_vs_conv <- lm(data = taiwan_data,
                        price_twd_msq ~ n_convenience)
mdl_price_vs_conv
```

```
Call:
lm(formula = price_twd_msq ~ n_convenience, data = taiwan_data)

Coefficients:
  (Intercept)  n_convenience
        8.224          0.798
```

In other words:

price_twd_msq = 8.224 + 0.798 * n_convenience

Simple linear regression (cont.)

```
> summary(mdl_price_vs_conv)

Call:
lm(formula = price_twd_msq ~ n_convenience, data = taiwan_data)

Residuals:
    Min      1Q  Median      3Q     Max
-10.713  -2.221  -0.541   1.810  26.530

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)     8.2242     0.2850    28.9 <0.0000000000000002 ***
n_convenience   0.7981     0.0565    14.1 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.38 on 412 degrees of freedom
Multiple R-squared:  0.326,	Adjusted R-squared:  0.324
F-statistic:  199 on 1 and 412 DF,  p-value: <0.0000000000000002
```

The coefficient of n_convenience of 0.7981 indicates that for each additional unit increase in n_convenience, the expected value of price_twd_msq increases by 0.7981, all else being equal.

Simple linear regression (cont.)

```
> summary(mdl_price_vs_conv)

Call:
lm(formula = price_twd_msq ~ n_convenience, data = taiwan_data)

Residuals:
    Min      1Q  Median      3Q     Max
-10.713  -2.221  -0.541   1.810  26.530

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)     8.2242     0.2850    28.9 <0.0000000000000002 ***
n_convenience   0.7981     0.0565    14.1 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.38 on 412 degrees of freedom
Multiple R-squared:  0.326,	Adjusted R-squared:  0.324
F-statistic:  199 on 1 and 412 DF,  p-value: <0.0000000000000002
```

Extremely small p-value (p-value < 0.05) suggests that the number of convenience stores nearby is a ***statistically significant*** predictor of the price per square meter of real estate in Taiwan.

In other words, the result suggests that the coefficient for n_convenience is significantly different from zero and <u>we can reject the null hypothesis</u> that there is no relationship between the number of convenience stores nearby and the price per square meter of real estate.

Simple linear regression (cont.)

```
> summary(mdl_price_vs_conv)

Call:
lm(formula = price_twd_msq ~ n_convenience, data = taiwan_data)

Residuals:
    Min      1Q  Median      3Q     Max
-10.713  -2.221  -0.541   1.810  26.530

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept)     8.2242     0.2850    28.9 <0.0000000000000002 ***
n_convenience   0.7981     0.0565    14.1 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.38 on 412 degrees of freedom
Multiple R-squared:  0.326,	Adjusted R-squared:  0.324
F-statistic:  199 on 1 and 412 DF,  p-value: <0.0000000000000002
```

The R-squared value of 0.326 indicates that about 32.6% of the variability in price_twd_msq can be explained by the linear relationship with n_convenience.

The adjusted R-squared value of 0.324 is similar, but takes into account the number of predictors in the model.

Note, that a low R-squared value suggests that the model explains only a small proportion of the variance in the dependent variable

Simple linear regression (cont.)

```
> summary(mdl_price_vs_conv)

Call:
lm(formula = price_twd_msq ~ n_convenience, data = taiwan_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.713  -2.221  -0.541   1.810  26.530 

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)    
(Intercept)     8.2242     0.2850    28.9 <0.0000000000000002 ***
n_convenience   0.7981     0.0565    14.1 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.38 on 412 degrees of freedom
Multiple R-squared:  0.326,	Adjusted R-squared:  0.324 
F-statistic:  199 on 1 and 412 DF,  p-value: <0.0000000000000002
```

p-value of <0.0000000000000002 suggest that the overall model is statistically significant, meaning that the independent variable n_convenience has a significant impact on the dependent variable price_twd_msq.

Multiple linear regression

```r
#Run a linear regression of price_twd_msq vs. n_convenience and dist_to_mrt_m
mdl_price_vs_conv_dist <- lm(data = taiwan_data,
                             price_twd_msq ~ n_convenience + dist_to_mrt_m)
mdl_price_vs_conv_dist
```

```
Call:
lm(formula = price_twd_msq ~ n_convenience + dist_to_mrt_m, data = taiwan_data)

Coefficients:
  (Intercept)  n_convenience  dist_to_mrt_m
     11.83749        0.36236       -0.00169
```

In other words:

price_twd_msq = 11.83749 + 0.36236 * n_convenience -

- 0.00169 * dist_to_mrt_m

Multiple linear regression (cont.)

```
> summary(mdl_price_vs_conv_dist)

Call:
lm(formula = price_twd_msq ~ n_convenience + dist_to_mrt_m, data = taiwan_data)

Residuals:
    Min      1Q  Median      3Q     Max
-11.048  -1.774  -0.411   1.447  23.779

Coefficients:
               Estimate Std. Error t value               Pr(>|t|)
(Intercept)   11.837489   0.393194   30.11 < 0.0000000000000002 ***
n_convenience  0.362360   0.061291    5.91           0.0000000071 ***
dist_to_mrt_m -0.001688   0.000143  -11.80 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.93 on 411 degrees of freedom
Multiple R-squared:  0.497,	Adjusted R-squared:  0.494
F-statistic:  203 on 2 and 411 DF,  p-value: <0.0000000000000002
```

For every one unit increase in n_convenience, the predicted value of the dependent variable increases by 0.36, and for every one unit increase in dist_to_mrt_m, the predicted value of the dependent variable decreases by 0.0017.

Multiple linear regression (cont.)

```
> summary(mdl_price_vs_conv_dist)

Call:
lm(formula = price_twd_msq ~ n_convenience + dist_to_mrt_m, data = taiwan_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.048  -1.774  -0.411   1.447  23.779 

Coefficients:
               Estimate Std. Error t value             Pr(>|t|)    
(Intercept)   11.837489   0.393194   30.11 < 0.0000000000000002 ***
n_convenience  0.362360   0.061291    5.91         0.0000000071 ***
dist_to_mrt_m -0.001688   0.000143  -11.80 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.93 on 411 degrees of freedom
Multiple R-squared:  0.497,	Adjusted R-squared:  0.494 
F-statistic:  203 on 2 and 411 DF,  p-value: <0.0000000000000002
```

Both coefficients are statistically significant with a p-value less than 0.05.

Therefore, the model is a good fit for the data and can be used to predict the dependent variable based on the values of the independent variables.

Multiple linear regression (cont.)

```
> summary(mdl_price_vs_conv_dist)

Call:
lm(formula = price_twd_msq ~ n_convenience + dist_to_mrt_m, data = taiwan_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.048  -1.774  -0.411   1.447  23.779 

Coefficients:
               Estimate Std. Error t value            Pr(>|t|)    
(Intercept)   11.837489   0.393194   30.11 < 0.0000000000000002 ***
n_convenience  0.362360   0.061291    5.91        0.0000000071 ***
dist_to_mrt_m -0.001688   0.000143  -11.80 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.93 on 411 degrees of freedom
Multiple R-squared:  0.497,	Adjusted R-squared:  0.494 
F-statistic:   203 on 2 and 411 DF,  p-value: <0.0000000000000002
```

The model is significant, as indicated by the low p-value and the multiple R-squared value of 0.497.

This suggests that the model explains a moderate proportion of the variance in the dependent variable.

Making a prediction

```r
# create a new data frame with values for the predictors
new_data <- data.frame(n_convenience = 5,
                       dist_to_mrt_m = 300)

# use the predict() function to make a prediction for the new data
prediction <- predict(mdl_price_vs_conv_dist, newdata = new_data)

# print the prediction
print(prediction)
```

```
> print(prediction)
    1
13.14
```

Which means that:

- for property which has 5 convenience stores located nearby and which is located in a distance of 300 m to the nearest MRT station, the price of the property should be 13.14 New Taiwan Dollars per square meter.

9. In-class Assignment

The data set “**mtcars**” you’ll be working with contains information on fuel consumption and performance of various car models. The dataset contains 11 columns:

mpg: Miles/(US) gallon

cyl: Number of cylinders

disp: Displacement (cu.in.)

hp: Gross horsepower

drat: Rear axle ratio

wt: Weight (lb/1000)

qsec: 1/4 mile time

vs: V/S (0 = V-shaped, 1 = straight)

am: Transmission (0 = automatic, 1 = manual)

gear: Number of forward gears

carb: Number of carburetors

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```