MAP 2112 – Introdução à Lógica de Programação e Modelagem Computacional

1º Semestre - 2020

Prof. Dr. Luis Carlos de Castro Santos

lsantos@ime.usp.br

Leve Introdução a Estatística com R

Leve Introdução a Visualização com R

Projeto de Data Science Individual

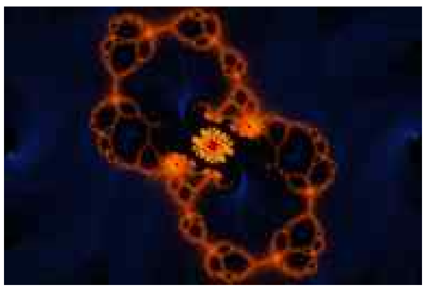
Projeto de Data Science em Grupo

R Tutorial

An R Introduction to Statistics

HOME DOWNLOAD SALES

Elementary Statistics with R



Ever wonder how to finish your statistics homework real fast? Or you just want a quick way to verify your tedious calculations in your statistics class assignment. We provide an answer here by solving statistics exercises with R.

Here, you will find statistics problems similar to those found in popular college textbooks. The R solutions are short, self-contained and requires minimal R skill. Most of them are just a few lines in length. With simple modifications, the code samples can be turned into homework answers. In additional to helping with your homework, the tutorials will give you a taste of working with statistics software in general, and it will prove invaluable in the success of your career.

We have included separate introductory tutorials for basic R concepts. The topics are by no means comprehensive. Nevertheless, even if you are not familiar with R, you can go through just the first *R Introduction* page. Then go straight to the statistics tutorials, and only come back for reference as needed.

http://www.r-tutor.com/elementary-statistics

O objetivo desse material é auxiliar na realização dos trabalhos. Não haverá cobrança em prova.

Elementary Statistics with R

- Qualitative Data
- Quantitative Data
- Numerical Measures
- Probability Distributions
- Interval Estimation
- Hypothesis Testing
- Type II Error
- Inference About Two Populations
- Goodness of Fit
- Analysis of Variance
- Non-parametric Methods
- Simple Linear Regression
- Multiple Linear Regression
- Logistic Regression

O objetivo desse material é auxiliar na realização dos trabalhos. Não haverá cobrança em prova.

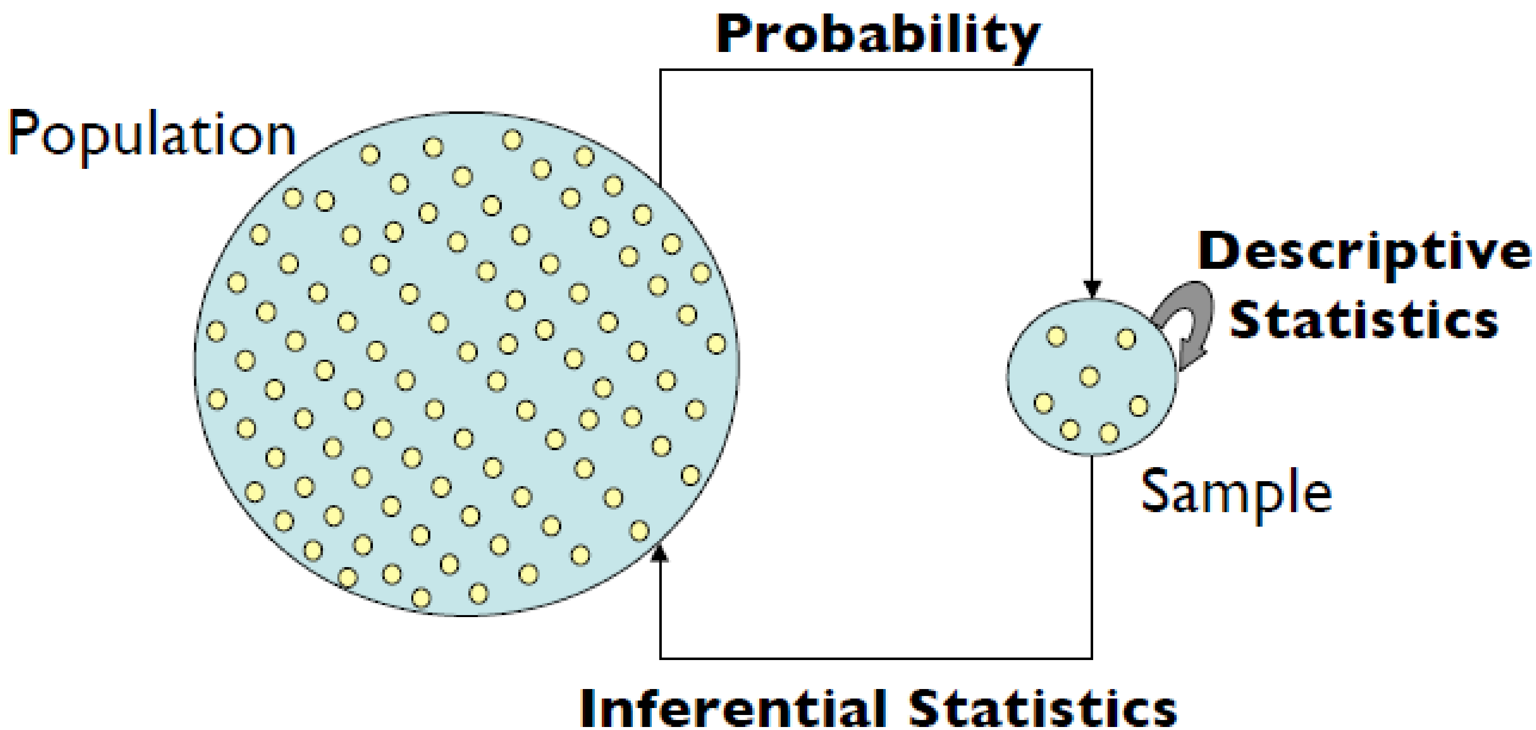
Probability Distributions

A **probability distribution** describes how the values of a random variable is distributed. For example, the collection of all possible outcomes of a sequence of coin tossing is known to follow the binomial distribution. Whereas the means of sufficiently large samples of a data population are known to resemble the normal distribution. Since the characteristics of these theoretical distributions are well understood, they can be used to make statistical inferences on the entire data population as a whole.

In the following tutorials, we demonstrate how to compute a few well-known probability distributions that occurs frequently in statistical study. We reference them quite often in other sections.

- Binomial Distribution
- Poisson Distribution
- Continuous Uniform Distribution
- Exponential Distribution
- Normal Distribution
- Chi-squared Distribution
- Student t Distribution
- F Distribution

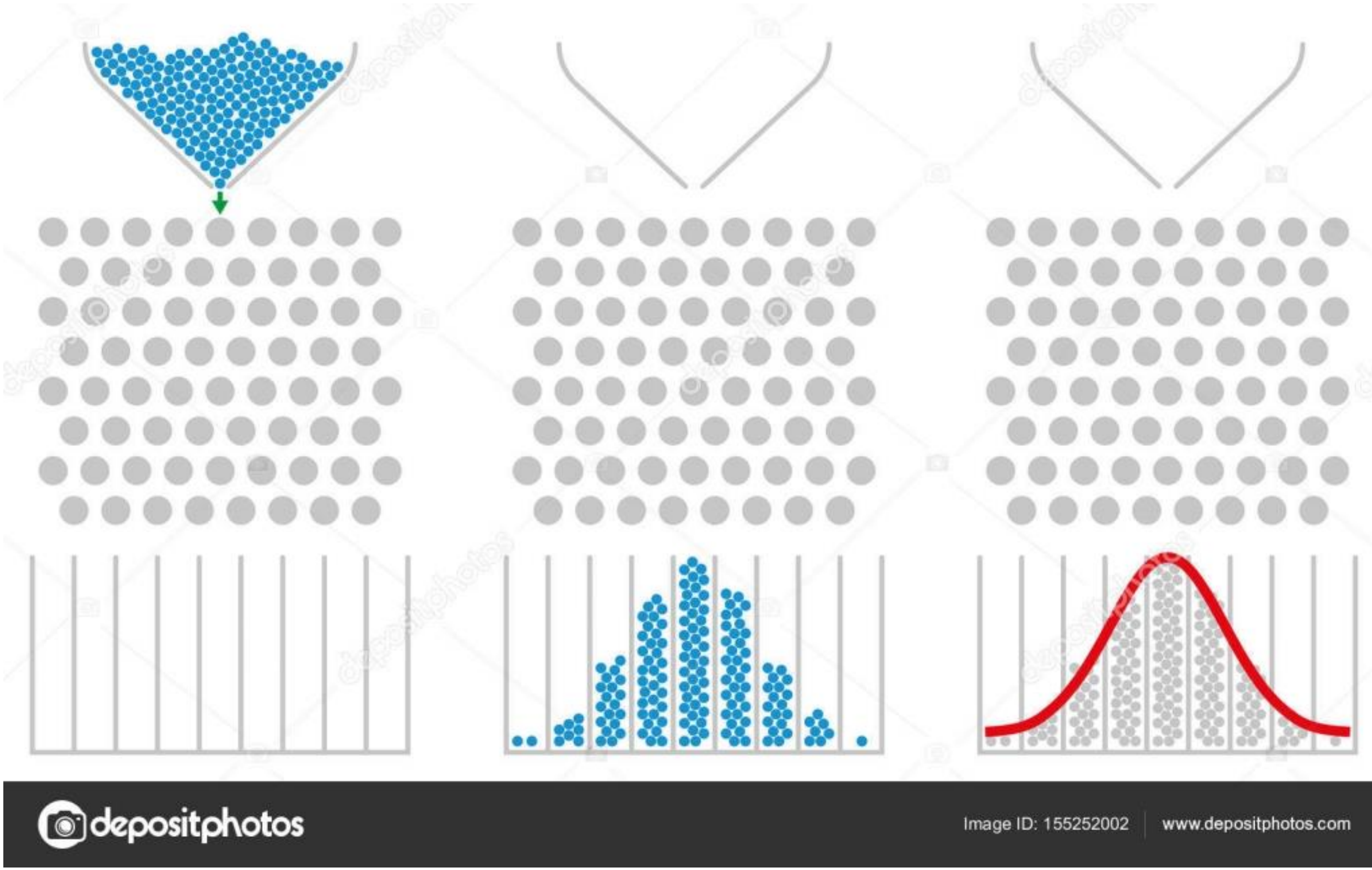
"Central Dogma" of Statistics



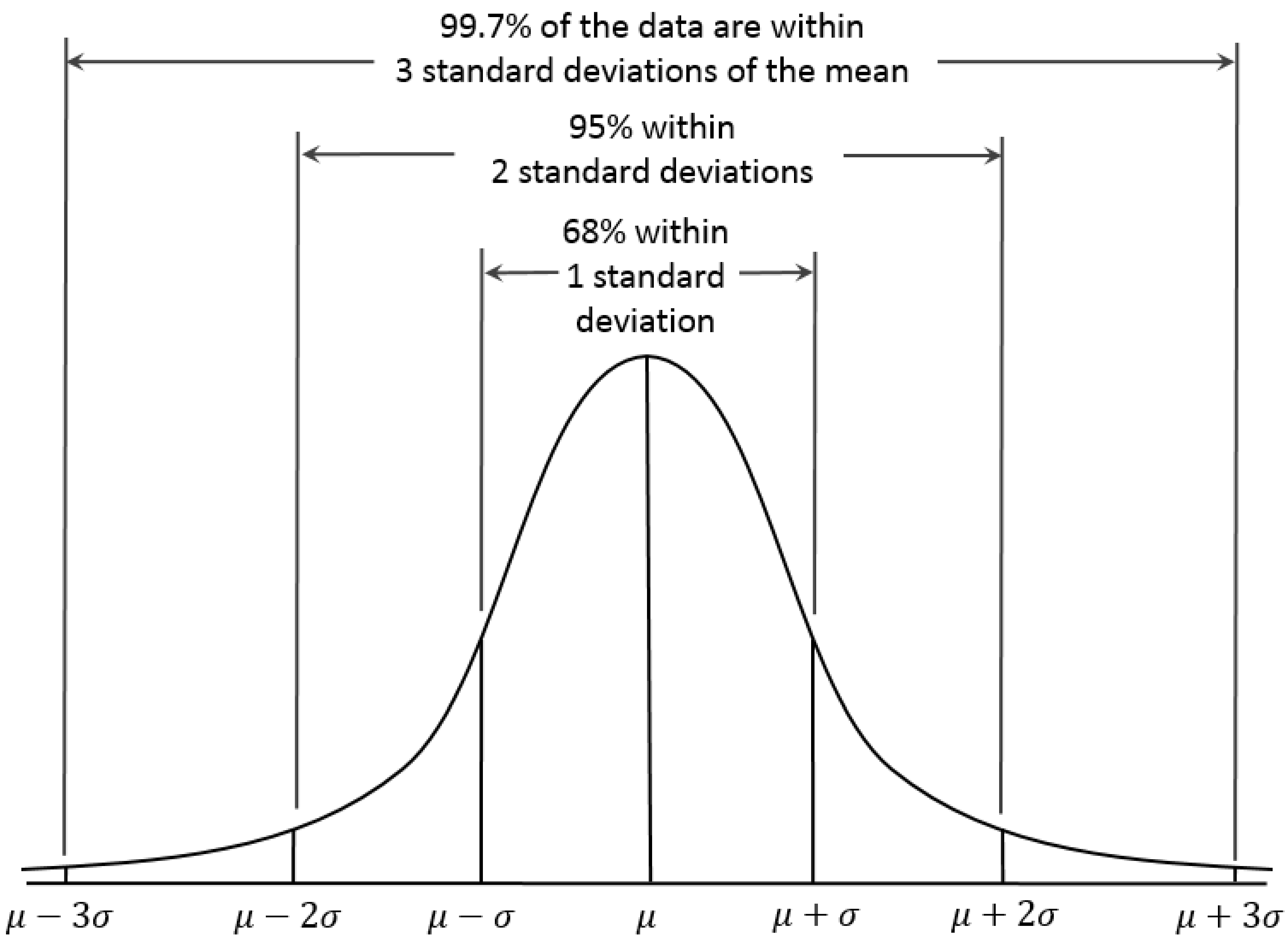
A inferência estatística consiste em inferior uma população a partir de uma amostra dessa população

Teorema do limite central

- Quanto maior for o tamanho *n* da amostra, mais a média amostral se aproximará da média da população.
- As propriedades da distribuição amostral asseguram que a média de uma amostra é uma boa estatística para inferir sobre a média da população μ da qual foi extraída.
- Ao mesmo tempo, o teorema do limite central estabelece que se o tamanho da amostra *n* for suficientemente grande a distribuição da média amostral será normal, qualquer que seja a forma da distribuição da população.
- Portanto, o teorema do limite central permite aplicar a distribuição normal para obter respostas da média de uma amostra de tamanho suficientemente grande retirada de uma população qualquer.



CLT: samples of observations of random variables independently drawn from independent distributions converge in distribution to the normal



99.7% of the data are within 3 standard deviations of the mean
95% within 2 standard deviations
68% within 1 standard deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Normal Distribution

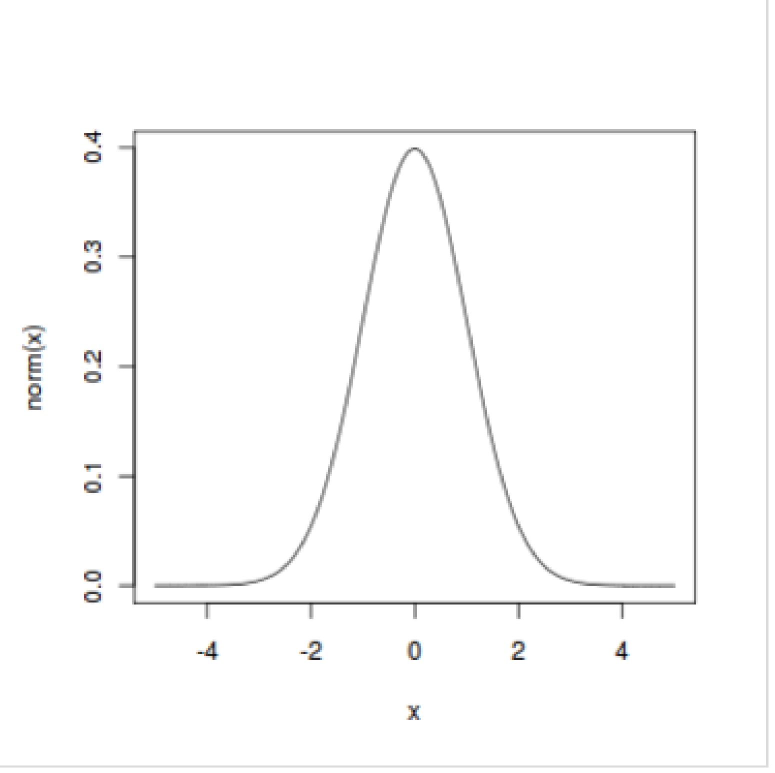
The **normal distribution** is defined by the following probability density function, where μ is the population mean and σ^2 is the variance.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

If a random variable X follows the normal distribution, then we write:

$$X \sim N(\mu, \sigma^2)$$

In particular, the normal distribution with $\mu = 0$ and $\sigma = 1$ is called the *standard normal distribution*, and is denoted as $N(0,1)$. It can be graphed as follows.



Problem

Assume that the test scores of a college entrance exam fits a normal distribution. Furthermore, the mean test score is 72, and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam?

Solution

We apply the function pnorm of the normal distribution with mean 72 and standard deviation 15.2. Since we are looking for the percentage of students scoring higher than 84, we are interested in the *upper tail* of the normal distribution.

```
> pnorm(84, mean=72, sd=15.2, lower.tail=FALSE)
[1] 0.21492
```

Answer

The percentage of students scoring 84 or more in the college entrance exam is 21.5%.

Simple Linear Regression

A **simple linear regression model** that describes the relationship between two variables x and y can be expressed by the following equation. The numbers α and β are called **parameters**, and ϵ is the **error term**.

$$y = \alpha + \beta x + \epsilon$$

For example, in the data set faithful, it contains sample data of two random variables named waiting and eruptions. The waiting variable denotes the waiting time until the next eruptions, and eruptions denotes the duration. Its linear regression model can be expressed as:

$$Eruptions = \alpha + \beta * Waiting + \epsilon$$

- Estimated Simple Regression Equation
- Coefficient of Determination
- Significance Test for Linear Regression
- Confidence Interval for Linear Regression
- Prediction Interval for Linear Regression
- Residual Plot
- Standardized Residual
- Normal Probability Plot of Residuals

Estimated Simple Regression Equation

If we choose the parameters α and β in the simple linear regression model so as to minimize the sum of squares of the error term ϵ, we will have the so called **estimated simple regression equation**. It allows us to compute **fitted values** of y based on values of x.

$$\hat{y} = a + bx$$

Problem

Apply the simple linear regression model for the data set faithful, and estimate the next eruption duration if the waiting time since the last eruption has been 80 minutes.

Solution

We apply the lm function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable eruption.lm.

```
> eruption.lm = lm(eruptions ~ waiting, data=faithful)
```

Then we extract the parameters of the estimated regression equation with the coefficients function.

```
> coeffs = coefficients(eruption.lm); coeffs
(Intercept)     waiting
  -1.874016    0.075628
```

We now fit the eruption duration using the estimated regression equation.

```
> waiting = 80           # the waiting time
> duration = coeffs[1] + coeffs[2]*waiting
> duration
(Intercept)
     4.1762
```

Answer

Based on the simple linear regression model, if the waiting time since the last eruption has been 80 minutes, we expect the next one to last 4.1762 minutes.

Alternative Solution

We wrap the waiting parameter value inside a new data frame named newdata.

```
> newdata = data.frame(waiting=80) # wrap the parameter
```

Then we apply the predict function to eruption.lm along with newdata.

```
> predict(eruption.lm, newdata)    # apply predict
     1
4.1762
```

Coefficient of Determination

The **coefficient of determination** of a linear regression model is the quotient of the variances of the fitted values and observed values of the dependent variable. If we denote y_i as the observed values of the dependent variable, $\bar{y}$ as its mean, and $\hat{y}_i$ as the fitted value, then the coefficient of determination is:

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Problem

Find the coefficient of determination for the simple linear regression model of the data set faithful.

Solution

We apply the lm function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable eruption.lm.

```
> eruption.lm = lm(eruptions ~ waiting, data=faithful)
```

Then we extract the coefficient of determination from the r.squared attribute of its summary.

```
> summary(eruption.lm)$r.squared
[1] 0.81146
```

Answer

The coefficient of determination of the simple linear regression model for the data set faithful is 0.81146.

Significance Test for Linear Regression

Assume that the error term ϵ in the linear regression model is independent of x, and is normally distributed, with zero mean and constant variance. We can decide whether there is any **significant relationship** between x and y by testing the null hypothesis that $\beta = 0$.

Problem

Decide whether there is a significant relationship between the variables in the linear regression model of the data set faithful at .05 significance level.

Solution

We apply the lm function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable eruption.lm.

```
> eruption.lm = lm(eruptions ~ waiting, data=faithful)
```

Then we print out the F-statistics of the significance test with the summary function.

```
> summary(eruption.lm)

Call:
lm(formula = eruptions ~ waiting, data = faithful)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2992 -0.3769  0.0351  0.3491  1.1933

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.87402    0.16014   -11.7   <2e-16 ***
waiting      0.07563    0.00222    34.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.497 on 270 degrees of freedom
Multiple R-squared: 0.811,      Adjusted R-squared: 0.811
F-statistic: 1.16e+03 on 1 and 270 DF,  p-value: <2e-16
```

Answer

As the p-value is much less than 0.05, we reject the null hypothesis that $\beta = 0$. Hence there is a significant relationship between the variables in the linear regression model of the data set faithful.

Confidence Interval for Linear Regression

Assume that the error term ϵ in the linear regression model is independent of x, and is normally distributed, with zero mean and constant variance. For a given value of x, the interval estimate for the mean of the dependent variable, $\bar{y}$, is called the **confidence interval**.

Problem

In the data set faithful, develop a 95% confidence interval of the mean eruption duration for the waiting time of 80 minutes.

Solution

We apply the lm function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable eruption.lm.

```
> attach(faithful)     # attach the data frame
> eruption.lm = lm(eruptions ~ waiting)
```

Then we create a new data frame that set the waiting time value.

```
> newdata = data.frame(waiting=80)
```

We now apply the predict function and set the predictor variable in the newdata argument. We also set the interval type as "confidence", and use the default 0.95 confidence level.

```
> predict(eruption.lm, newdata, interval="confidence")
     fit    lwr    upr
1 4.1762 4.1048 4.2476
> detach(faithful)     # clean up
```

Answer

The 95% confidence interval of the mean eruption duration for the waiting time of 80 minutes is between 4.1048 and 4.2476 minutes.

Prediction Interval for Linear Regression

Assume that the error term ϵ in the simple linear regression model is independent of x, and is normally distributed, with zero mean and constant variance. For a given value of x, the interval estimate of the dependent variable y is called the **prediction interval**.

Problem

In the data set faithful, develop a 95% prediction interval of the eruption duration for the waiting time of 80 minutes.

Solution

We apply the lm function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable eruption.lm.

```
> attach(faithful)     # attach the data frame
> eruption.lm = lm(eruptions ~ waiting)
```

Then we create a new data frame that set the waiting time value.

```
> newdata = data.frame(waiting=80)
```

We now apply the predict function and set the predictor variable in the newdata argument. We also set the interval type as "predict", and use the default 0.95 confidence level.

```
> predict(eruption.lm, newdata, interval="predict")
     fit    lwr    upr
1 4.1762 3.1961 5.1564
> detach(faithful)     # clean up
```

Answer

The 95% prediction interval of the eruption duration for the waiting time of 80 minutes is between 3.1961 and 5.1564 minutes.

Residual Plot

The **residual** data of the simple linear regression model is the difference between the observed data of the dependent variable y and the fitted values $\hat{y}$.

$$Residual = y - \hat{y}$$

Problem

Plot the residual of the simple linear regression model of the data set faithful against the independent variable waiting.

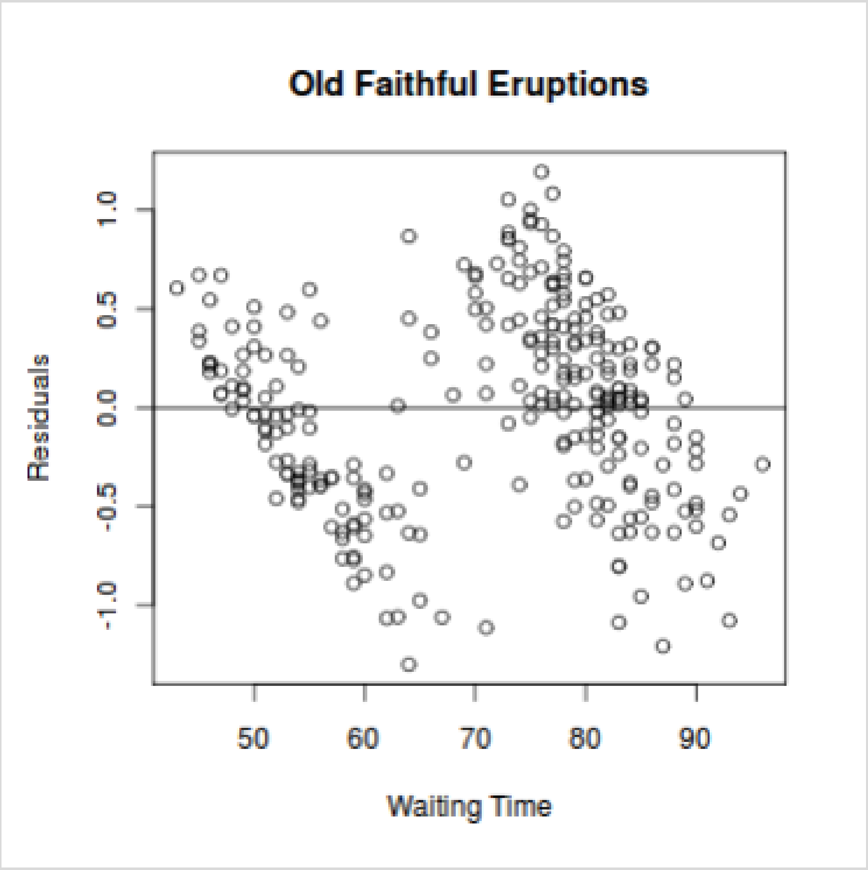
Solution

We apply the lm function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable eruption.lm. Then we compute the residual with the resid function.

```
> eruption.lm = lm(eruptions ~ waiting, data=faithful)
> eruption.res = resid(eruption.lm)
```

We now plot the residual against the observed values of the variable waiting.

```
> plot(faithful$waiting, eruption.res,
+     ylab="Residuals", xlab="Waiting Time",
+     main="Old Faithful Eruptions")
> abline(0, 0)                  # the horizon
```



Standardized Residual

The **standardized residual** is the residual divided by its standard deviation.

$$Standardized\ Residual\ i = \frac{Residual\ i}{Standard\ Deviation\ of\ Residual\ i}$$

Problem

Plot the standardized residual of the simple linear regression model of the data set faithful against the independent variable waiting.

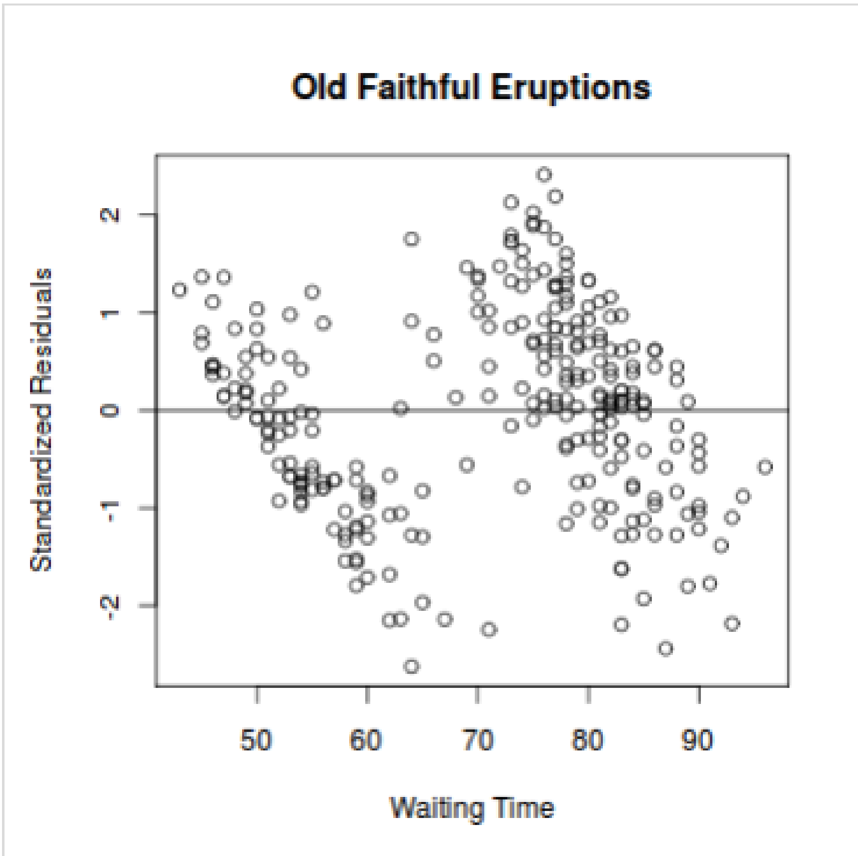
Solution

We apply the lm function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable eruption.lm. Then we compute the standardized residual with the rstandard function.

```
> eruption.lm = lm(eruptions ~ waiting, data=faithful)
> eruption.stdres = rstandard(eruption.lm)
```

We now plot the standardized residual against the observed values of the variable waiting.

```
> plot(faithful$waiting, eruption.stdres,
+     ylab="Standardized Residuals",
+     xlab="Waiting Time",
+     main="Old Faithful Eruptions")
> abline(0, 0)                  # the horizon
```



Normal Probability Plot of Residuals

The **normal probability plot** is a graphical tool for comparing a data set with the normal distribution. We can use it with the standardized residual of the linear regression model and see if the error term ϵ is actually normally distributed.

Problem

Create the normal probability plot for the standardized residual of the data set faithful.

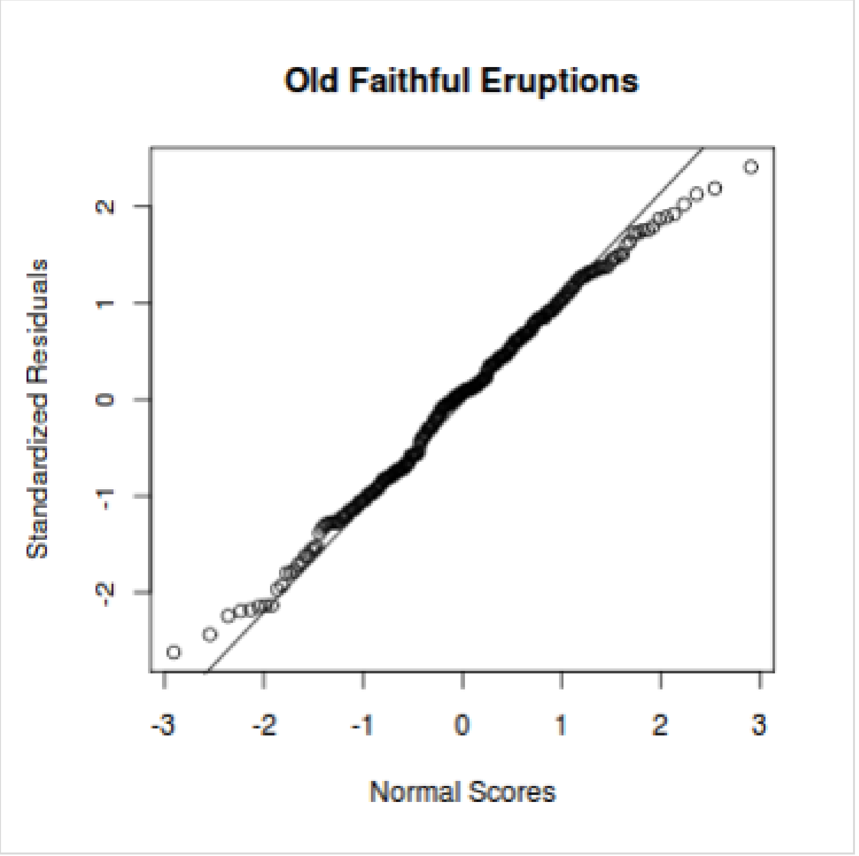
Solution

We apply the lm function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable eruption.lm. Then we compute the standardized residual with the rstandard function.

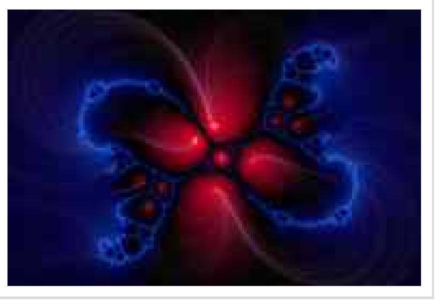
```
> eruption.lm = lm(eruptions ~ waiting, data=faithful)
> eruption.stdres = rstandard(eruption.lm)
```

We now create the normal probability plot with the qqnorm function, and add the qqline for further comparison.

```
> qqnorm(eruption.stdres,
+     ylab="Standardized Residuals",
+     xlab="Normal Scores",
+     main="Old Faithful Eruptions")
> qqline(eruption.stdres)
```



Multiple Linear Regression



A **multiple linear regression** (MLR) model that describes a dependent variable y by independent variables $x_1, x_2, \ldots, x_p$ ($p > 1$) is expressed by the equation as follows, where the numbers α and β_k ($k = 1, 2, \ldots, p$) are the **parameters**, and ϵ is the **error term**.

$$y = \alpha + \sum_k \beta_k x_k + \epsilon$$

For example, in the built-in data set stackloss from observations of a chemical plant operation, if we assign stackloss as the dependent variable, and assign Air.Flow (cooling air flow), Water.Temp (inlet water temperature) and Acid.Conc. (acid concentration) as independent variables, the multiple linear regression model is:

$$Stack.Loss = \alpha + \beta_1 * Air.Flow + \beta_2 * Water.Temp + \beta_3 * Acid.Conc. + \epsilon$$

Further detail of the stackloss data set can be found in the R documentation.

```
> help(stackloss)
```

- Estimated Multiple Regression Equation
- Multiple Coefficient of Determination
- Adjusted Coefficient of Determination
- Significance Test for MLR
- Confidence Interval for MLR
- Prediction Interval for MLR

Estimated Multiple Regression Equation

If we choose the parameters α and β_k (k = 1, 2, ..., p) in the multiple linear regression model so as to minimize the sum of squares of the error term ϵ, we will have the so called **estimated multiple regression equation.** It allows us to compute **fitted values** of y based on a set of values of x_k (k = 1, 2, ..., p) .

$$\hat{y} = a + \sum_k b_k x_k$$

Problem

Apply the multiple linear regression model for the data set stackloss, and predict the stack loss if the air flow is 72, water temperature is 20 and acid concentration is 85.

Solution

We apply the lm function to a formula that describes the variable stack.loss by the variables Air.Flow, Water.Temp and Acid.Conc. And we save the linear regression model in a new variable stackloss.lm.

```
> stackloss.lm = lm(stack.loss ~
+     Air.Flow + Water.Temp + Acid.Conc.,
+     data=stackloss)
```

We also wrap the parameters inside a new data frame named newdata.

```
> newdata = data.frame(Air.Flow=72,  # wrap the parameters
+     Water.Temp=20,
+     Acid.Conc.=85)
```

Lastly, we apply the predict function to stackloss.lm and newdata.

```
> predict(stackloss.lm, newdata)
     1
24.582
```

Answer

Based on the multiple linear regression model and the given parameters, the predicted stack loss is 24.582.

Multiple Coefficient of Determination

The **coefficient of determination** of a multiple linear regression model is the quotient of the variances of the fitted values and observed values of the dependent variable. If we denote y_i as the observed values of the dependent variable, $\bar{y}$ as its mean, and $\hat{y}_i$ as the fitted value, then the coefficient of determination is:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Problem

Find the coefficient of determination for the multiple linear regression model of the data set stackloss.

Solution

We apply the lm function to a formula that describes the variable stack.loss by the variables Air.Flow, Water.Temp and Acid.Conc. And we save the linear regression model in a new variable stackloss.lm.

```
> stackloss.lm = lm(stack.loss ~
+     Air.Flow + Water.Temp + Acid.Conc.,
+     data=stackloss)
```

Then we extract the coefficient of determination from the r.squared attribute of its summary.

```
> summary(stackloss.lm)$r.squared
[1] 0.91358
```

Answer

The coefficient of determination of the multiple linear regression model for the data set stackloss is 0.91358.

Adjusted Coefficient of Determination

The **adjusted coefficient of determination** of a multiple linear regression model is defined in terms of the coefficient of determination as follows, where n is the number of observations in the data set, and p is the number of independent variables.

$$R_{adj}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Problem

Find the adjusted coefficient of determination for the multiple linear regression model of the data set stackloss.

Solution

We apply the lm function to a formula that describes the variable stack.loss by the variables Air.Flow, Water.Temp and Acid.Conc. And we save the linear regression model in a new variable stackloss.lm.

```
> stackloss.lm = lm(stack.loss ~
+     Air.Flow + Water.Temp + Acid.Conc.,
+     data=stackloss)
```

Then we extract the coefficient of determination from the adj.r.squared attribute of its summary.

```
> summary(stackloss.lm)$adj.r.squared
[1]  0.89833
```

Answer

The adjusted coefficient of determination of the multiple linear regression model for the data set stackloss is 0.89833.

Significance Test for MLR

Assume that the error term ϵ in the multiple linear regression (MLR) model is independent of x_k (k = 1, 2, ..., p), and is normally distributed, with zero mean and constant variance. We can decide whether there is any **significant relationship** between the dependent variable y and any of the independent variables x_k (k = 1, 2, ..., p).

Problem

Decide which of the independent variables in the multiple linear regression model of the data set stackloss are statistically significant at .05 significance level.

Solution

We apply the lm function to a formula that describes the variable stack.loss by the variables Air.Flow, Water.Temp and Acid.Conc. And we save the linear regression model in a new variable stackloss.lm.

```
> stackloss.lm = lm(stack.loss ~
+     Air.Flow + Water.Temp + Acid.Conc.,
+     data=stackloss)
```

The t values of the independent variables can be found with the summary function.

```
> summary(stackloss.lm)

Call:
lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
    data = stackloss)

Residuals:
   Min     1Q Median     3Q    Max
-7.238 -1.712 -0.455  2.361  5.698

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -39.920     11.896   -3.36   0.0038 **
Air.Flow       0.716      0.135    5.31  5.8e-05 ***
Water.Temp     1.295      0.368    3.52   0.0026 **
Acid.Conc.    -0.152      0.156   -0.97   0.3440
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.24 on 17 degrees of freedom
Multiple R-squared: 0.914,      Adjusted R-squared: 0.898
F-statistic: 59.9 on 3 and 17 DF,  p-value: 3.02e-09
```

Answer

As the p-values of Air.Flow and Water.Temp are less than 0.05, they are both statistically significant in the multiple linear regression model of stackloss.

Confidence Interval for MLR

Assume that the error term ϵ in the multiple linear regression (MLR) model is independent of x_k (k = 1, 2, ..., p), and is normally distributed, with zero mean and constant variance. For a given set of values of x_k (k = 1, 2, ..., p), the interval estimate for the mean of the dependent variable, $\bar{y}$, is called the **confidence interval**.

Problem

In data set stackloss, develop a 95% confidence interval of the stack loss if the air flow is 72, water temperature is 20 and acid concentration is 85.

Solution

We apply the lm function to a formula that describes the variable stack.loss by the variables Air.Flow, Water.Temp and Acid.Conc. And we save the linear regression model in a new variable stackloss.lm.

```
> attach(stackloss)    # attach the data frame
> stackloss.lm = lm(stack.loss ~
+     Air.Flow + Water.Temp + Acid.Conc.)
```

Then we wrap the parameters inside a new data frame variable newdata.

```
> newdata = data.frame(Air.Flow=72,
+     Water.Temp=20,
+     Acid.Conc.=85)
```

We now apply the predict function and set the predictor variable in the newdata argument. We also set the interval type as "confidence", and use the default 0.95 confidence level.

```
> predict(stackloss.lm, newdata, interval="confidence")
     fit    lwr    upr
1 24.582 20.218 28.945
> detach(stackloss)    # clean up
```

Answer

The 95% confidence interval of the stack loss with the given parameters is between 20.218 and 28.945.

Prediction Interval for MLR

Assume that the error term ϵ in the multiple linear regression (MLR) model is independent of x_k (k = 1, 2, ..., p), and is normally distributed, with zero mean and constant variance. For a given set of values of x_k (k = 1, 2, ..., p), the interval estimate of the dependent variable y is called the **prediction interval**.

Problem

In data set stackloss, develop a 95% prediction interval of the stack loss if the air flow is 72, water temperature is 20 and acid concentration is 85.

Solution

We apply the lm function to a formula that describes the variable stack.loss by the variables Air.Flow, Water.Temp and Acid.Conc. And we save the linear regression model in a new variable stackloss.lm.

```
> attach(stackloss)    # attach the data frame
> stackloss.lm = lm(stack.loss ~
+     Air.Flow + Water.Temp + Acid.Conc.)
```

Then we wrap the parameters inside a new data frame variable newdata.

```
> newdata = data.frame(Air.Flow=72,
+     Water.Temp=20,
+     Acid.Conc.=85)
```

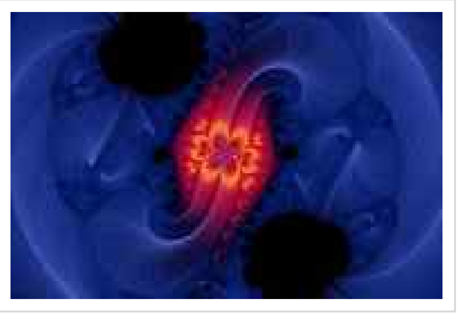
We now apply the predict function and set the predictor variable in the newdata argument. We also set the interval type as "predict", and use the default 0.95 confidence level.

```
> predict(stackloss.lm, newdata, interval="predict")
     fit    lwr    upr
1 24.582 16.466 32.697
> detach(stackloss)    # clean up
```

Answer

The 95% confidence interval of the stack loss with the given parameters is between 16.466 and 32.697.

Logistic Regression



We use the **logistic regression equation** to predict the probability of a dependent variable taking the dichotomy values 0 or 1. Suppose $x_1, x_2, \ldots, x_p$ are the independent variables, α and β_k ($k = 1, 2, \ldots, p$) are the parameters, and $E(y)$ is the expected value of the dependent variable y, then the logistic regression equation is:

$$E(y) = 1/(1 + e^{-(\alpha + \sum_k \beta_k x_k)})$$

For example, in the built-in data set mtcars, the data column am represents the transmission type of the automobile model (0 = automatic, 1 = manual). With the logistic regression equation, we can model the probability of a manual transmission in a vehicle based on its engine horsepower and weight data.

$$P(Manual\ Transmission) = 1/(1 + e^{-(\alpha + \beta_1 * Horsepower + \beta_2 * Weight)})$$

- Estimated Logistic Regression Equation
- Significance Test for Logistic Regression

Estimated Logistic Regression Equation

Using the generalized linear model, an **estimated logistic regression equation** can be formulated as below. The coefficients a and b_k (k = 1, 2, ..., p) are determined according to a maximum likelihood approach, and it allows us to estimate the probability of the dependent variable y taking on the value 1 for given values of x_k (k = 1, 2, ..., p).

$$Estimate\ of\ P(y = 1 \mid x_1, ...x_p) = 1/(1 + e^{-(a+\sum_k b_k x_k)})$$

Problem

By use of the logistic regression equation of vehicle transmission in the data set mtcars, estimate the probability of a vehicle being fitted with a manual transmission if it has a 120hp engine and weights 2800 lbs.

Solution

We apply the function glm to a formula that describes the transmission type (am) by the horsepower (hp) and weight (wt). This creates a generalized linear model (GLM) in the binomial family.

```
> am.glm = glm(formula=am ~ hp + wt,
+              data=mtcars,
+              family=binomial)
```

We then wrap the test parameters inside a data frame newdata.

```
> newdata = data.frame(hp=120, wt=2.8)
```

Now we apply the function predict to the generalized linear model am.glm along with newdata. We will have to select *response* prediction type in order to obtain the predicted probability.

```
> predict(am.glm, newdata, type="response")
      1
0.64181
```

Answer

For an automobile with 120hp engine and 2800 lbs weight, the probability of it being fitted with a manual transmission is about 64%.

Significance Test for Logistic Regression

We can decide whether there is any significant relationship between the dependent variable y and the independent variables x_k ($k = 1, 2, ..., p$) in the logistic regression equation. In particular, if any of the null hypothesis that $\beta_k = 0$ ($k = 1, 2, ..., p$) is valid, then x_k is statistically insignificant in the logistic regression model.

Problem

At .05 significance level, decide if any of the independent variables in the logistic regression model of vehicle transmission in data set mtcars is statistically insignificant.

Solution

We apply the function glm to a formula that describes the transmission type (am) by the horsepower (hp) and weight (wt). This creates a generalized linear model (GLM) in the binomial family.

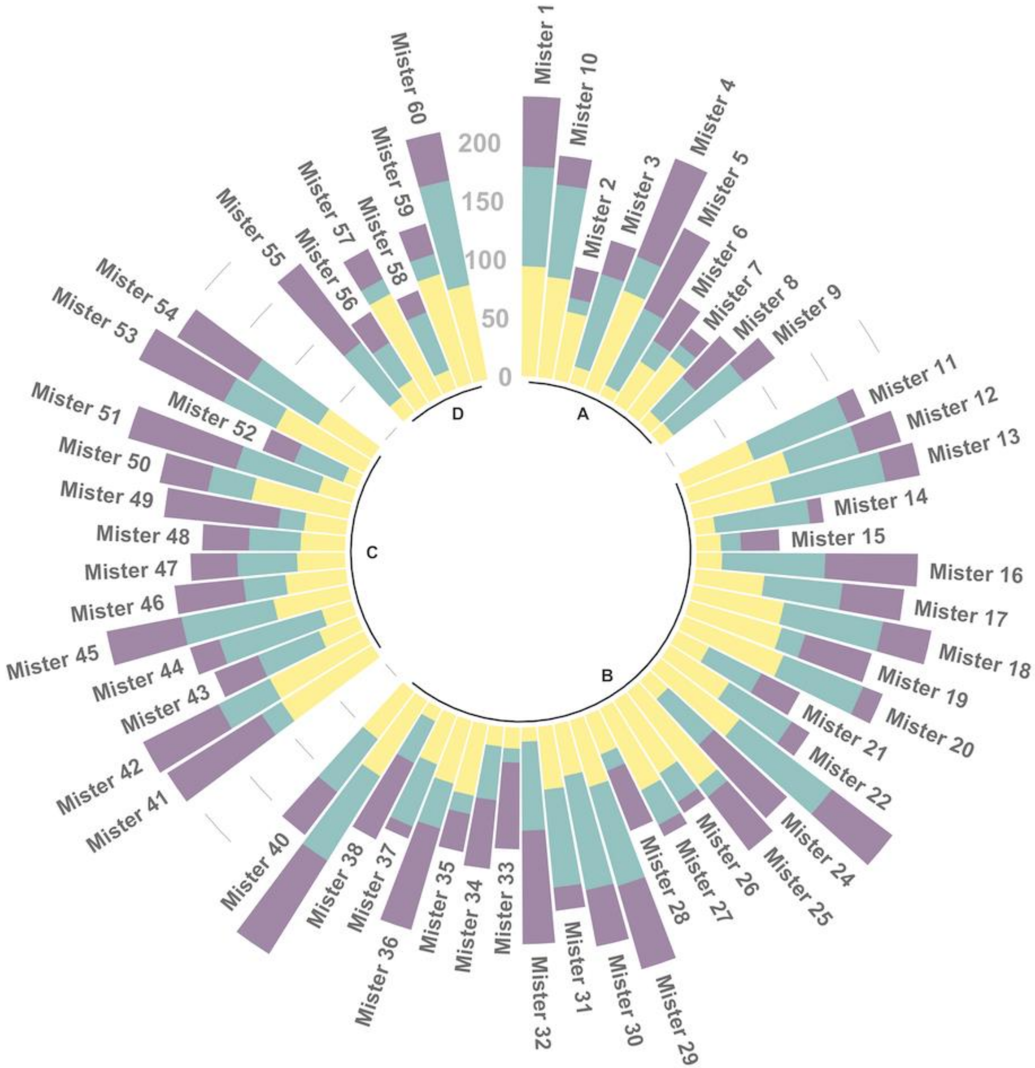
```
> am.glm = glm(formula=am ~ hp + wt,
+              data=mtcars,
+              family=binomial)
```

We then print out the summary of the generalized linear model and check for the p-values of the hp and wt variables.

```
> summary(am.glm)

Call:
glm(formula = am ~ hp + wt, family = binomial, data = mtcars)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2537  -0.1568  -0.0168   0.1543   1.3449

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  18.8663     7.4436    2.53   0.0113 *
hp            0.0363     0.0177    2.04   0.0409 *
wt           -8.0835     3.0687   -2.63   0.0084 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.230  on 31  degrees of freedom
Residual deviance: 10.059  on 29  degrees of freedom
AIC: 16.06

Number of Fisher Scoring iterations: 8
```

Answer

As the p-values of the hp and wt variables are both less than 0.05, neither hp or wt is insignificant in the logistic regression model.



https://www.r-graph-gallery.com/299-circular-stacked-barplot.html