



Statistical tools for high-throughput data analysis

Licence:








Search...

[Home](#)[Basics](#)[Data](#)[Visualize](#)[Analyze](#)[Products](#)[Contribute](#)[Support](#)[About](#)

[Home](#) / [Articles](#) / [Machine Learning](#) / [Regression Analysis](#) / Predict in R: Model Predictions and Confidence Intervals

Articles - Regression Analysis

Predict in R: Model Predictions and Confidence Intervals

 [kassambara](#) |  10/03/2018 |  7898 |  [Comments \(4\)](#) |  [Regression Analysis](#)

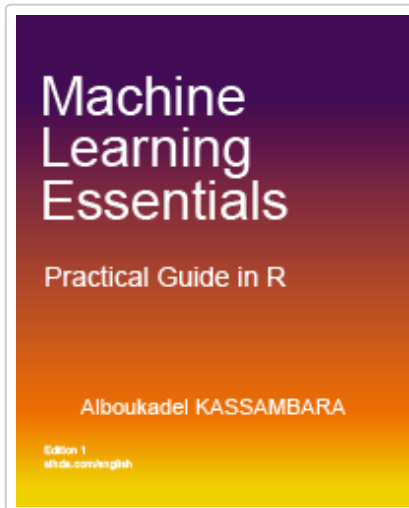
The main goal of **linear regression** is to **predict** an outcome value on the basis of one or multiple predictor variables.

In this chapter, we'll describe how to predict outcome for new observations data using R.. You will also learn how to display the confidence intervals and the prediction intervals.

Contents:

- [Build a linear regression](#)
- [Prediction for new data set](#)
- [Confidence interval](#)
- [Prediction interval](#)
- [Prediction interval or confidence interval?](#)
- [Discussion](#)
- [References](#)

The Book:



Machine Learning Essentials:
Practical Guide in R

Build a linear regression

We start by building a simple linear regression model that predicts the stopping distances of cars on the basis of the speed.

```
# Load the data
data("cars", package = "datasets")
# Build the model
model <- lm(dist ~ speed, data = cars)
model
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Coefficients:
## (Intercept)      speed
##      -17.58         3.93
```

The linear model equation can be written as follow: $\text{dist} = -17.579 + 3.932 \times \text{speed}$.

Note that, the units of the variable `speed` and `dist` are respectively, `mph` and `ft`.

Prediction for new data set

Using the above model, we can predict the stopping distance for a new speed value.

Start by creating a new data frame containing, for example, three new speed values:

```
new.speeds <- data.frame(
  speed = c(12, 19, 24)
```

```
)
```

You can predict the corresponding stopping distances using the R function `predict()` as follow:

```
predict(model, newdata = new.speeds)
```

```
##      1      2      3  
## 29.6 57.1 76.8
```

Confidence interval

The confidence interval reflects the uncertainty around the mean predictions. To display the 95% confidence intervals around the mean the predictions, specify the option `interval = "confidence"`:

```
predict(model, newdata = new.speeds, interval = "confidence")
```

```
##      fit   lwr   upr  
## 1 29.6 24.4 34.8  
## 2 57.1 51.8 62.4  
## 3 76.8 68.4 85.2
```

The output contains the following columns:

- `fit`: the predicted sale values for the three new advertising budget
- `lwr` and `upr`: the lower and the upper confidence limits for the expected values, respectively. By default the function produces the 95% confidence limits.

For example, the 95% confidence interval associated with a speed of 19 is (51.83, 62.44). This means that, according to our model, a car with a speed of 19 mph has, on average, a stopping distance ranging between 51.83 and 62.44 ft.

Prediction interval

The prediction interval gives uncertainty around a single value. In the same way, as the confidence intervals, the prediction intervals can be computed as follow:

```
predict(model, newdata = new.speeds, interval = "prediction")
```

```
##      fit   lwr   upr  
## 1 29.6 -1.75 61.0  
## 2 57.1 25.76 88.5  
## 3 76.8 44.75 108.8
```

The 95% prediction intervals associated with a speed of 19 is (25.76, 88.51). This means that, according to our model, 95% of the cars with a speed of 19 mph have a stopping distance between 25.76 and 88.51.



Note that, prediction interval relies strongly on the assumption that the residual errors are normally distributed with a constant variance. So, you should only use such intervals if you believe that the assumption is approximately met for the data at hand.

Prediction interval or confidence interval?

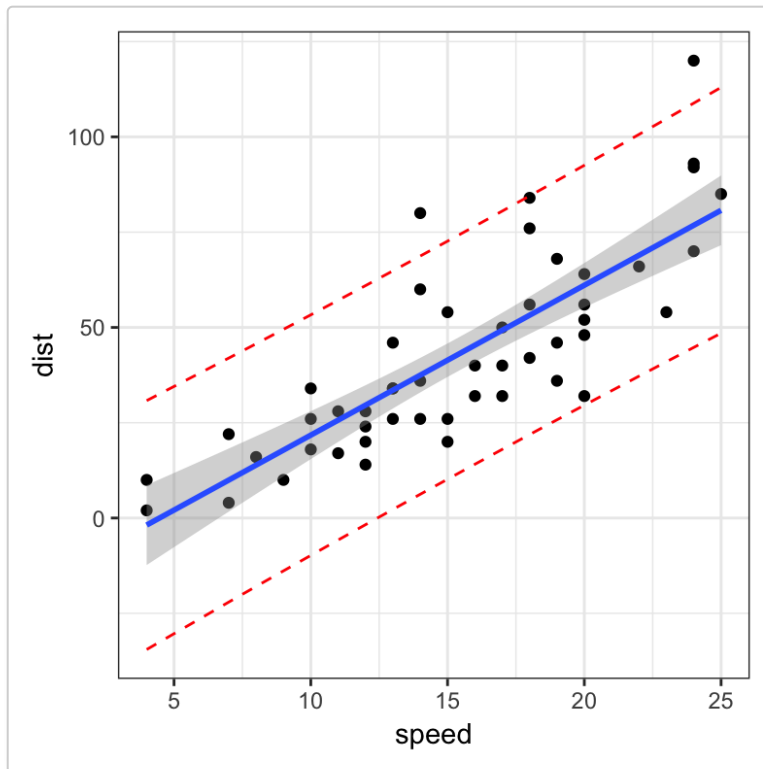
A prediction interval reflects the uncertainty around a single value, while a confidence interval reflects the uncertainty around the mean prediction values. Thus, a prediction interval will be generally much wider than a confidence interval for the same value.

Which one should we use? The answer to this question depends on the context and the purpose of the analysis. Generally, we are interested in specific individual predictions, so a prediction interval would be more appropriate. Using a confidence interval when you should be using a prediction interval will greatly underestimate the uncertainty in a given predicted value (P. Bruce and Bruce 2017).

The R code below creates a scatter plot with:

- The regression line in blue
- The confidence band in gray
- The prediction band in red

```
# 0. Build linear model
data("cars", package = "datasets")
model <- lm(dist ~ speed, data = cars)
# 1. Add predictions
pred.int <- predict(model, interval = "prediction")
mydata <- cbind(cars, pred.int)
# 2. Regression line + confidence intervals
library("ggplot2")
p <- ggplot(mydata, aes(speed, dist)) +
  geom_point() +
  stat_smooth(method = lm)
# 3. Add prediction intervals
p + geom_line(aes(y = lwr), color = "red", linetype = "dashed")+
  geom_line(aes(y = upr), color = "red", linetype = "dashed")
```



Discussion

In this chapter, we have described how to use the R function `predict()` for predicting outcome for new data.

References

Bruce, Peter, and Andrew Bruce. 2017. *Practical Statistics for Data Scientists*. O'Reilly Media.

Last update : 24/07/2018

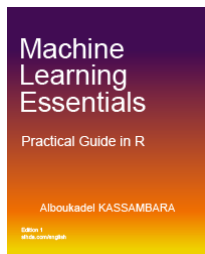
★★★★☆ 1 Note



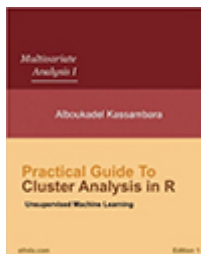
Enjoyed this article? Give us 5 stars ★★★★★ (just above this text block)! Reader needs to be STH-DA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

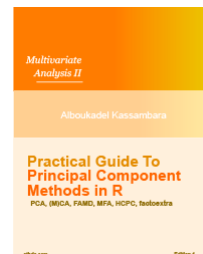
Recommended for You!



Machine Learning Essentials:
Practical Guide in R



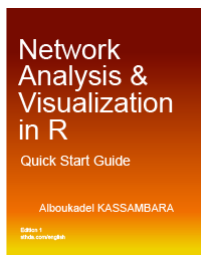
Practical Guide to Cluster Analysis in R



Practical Guide to Principal Component Methods in R



R Graphics Essentials for Great Data Visualization



Network Analysis and Visualization in R



More books on R and data science

The fields marked with a * are required !

Add a comment

Name

Visitor

Message

☺ **B** *I* U ~~S~~ 🔍 T! A 📄 📋 ☐ 🗨️ 🔍 ⚠️ 🌐 🖼️ 📷


Preview

* Code de vérification

How many vowels are in the word sthda?



kassambara 07/24/2018 at 22h02

Administrator

Fixed now, thank you @genghiskhan!

#565



genghiskhan 07/18/2018 at 01h17

Member

Thanks for your tutorial.

I think this equation should have the plus sign rather than minus.

$\text{dist} = -17.579 - 3.932 * \text{speed}$

It should be $\text{dist} = -17.579 + 3.932 * \text{speed}$

#559



kassambara 05/22/2018 at 22h47

Administrator

Thank you! Updated know

#492



Raul 05/22/2018 at 22h18

Visitor

nice article. one detail, when it says "a stopping distance ranging between 51.83 and 62.44 mph", it should say "a stopping distance ranging between 51.83 and 62.44 ft"

#491

Sign in

Login

Password

Auto connect

[Register](#)[Forgotten password](#)

Welcome!

Want to Learn More on R Programming and Data Science?

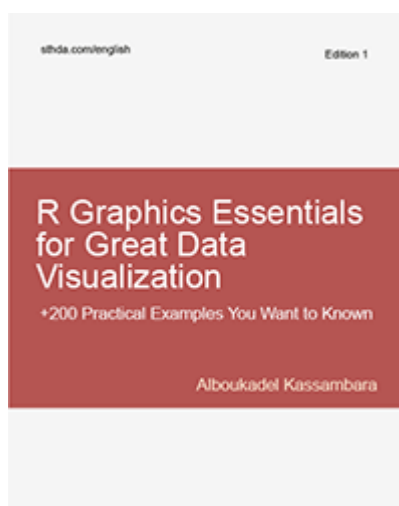
Follow us [by Email](#)

by [FeedBurner](#)

factoextra **survminer** **ggpubr**

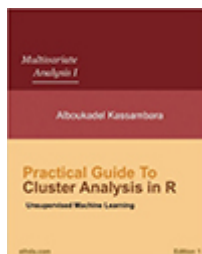
 **ggcorrplot** **fastqcr**

Our Books

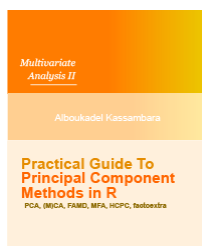


R Graphics Essentials for Great Data Visualization: 200 Practical Examples You Want to Know for Data Science

★ NEW!!

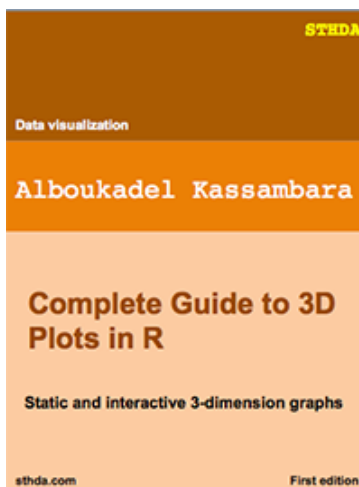


Practical Guide to Cluster Analysis in R



Practical Guide to Principal Component Methods in R

3D Plots in R



Guest Book

I'm psychologist, from Chile. This website is WONDERFUL!! Comprehensive, clear, simple, great!!!!

Thank you, thank you!!!!

Pablo

By *Visitor*

[Guest Book](#)

 **R-Bloggers**

Newsletter

Email



Boosted by PHPBoost
