Machine Learning with R: Linear Regression

Shaila Jamal

Data Analysis Support Assistant, DASH, McMaster Library

Ph.D. Candidate, School of Earth, Environment and Society, McMaster University

October 21, 2022

McMaster University | DASH

McMaster University | Library

McMaster University sits on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the "Dish With One Spoon" wampum agreement.

Session Recording and Privacy

This session is being recorded with the intention of being shared publicly via the web for future audiences.

In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.

Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.

Code of Conduct

The DASH program and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information:

scds.ca/events/code-of-conduct/

Certificate Program

The Sherman Centre offers a Certificate of Completion that rewards synchronous participation at 7 workshops. We also offer concentrations in Data Analysis and Visualization, Digital Scholarship, and Research Data Management.

Learn more about the Certificate Program: https://scds.ca/certificate-program

If you would like to be considered for the certificate, verify your participation in this form: https://u.mcmaster.ca/verification

At an unspecified point during the workshop, a code will be read aloud. This is the answer to the third question of the form.

Linear Regression

“A linear regression is a statistical model that analyzes the relationship between a response variable (often called y) and one or more variables and their interactions (often called x or explanatory variables).”

- “For example, when you calculate the age of a child based on their height, you are assuming the older they are, the taller they will be.”

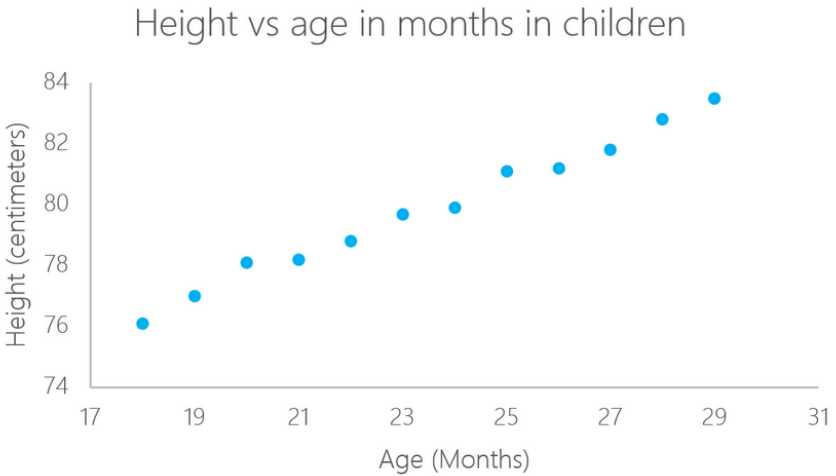
- “Not every problem can be solved with the same algorithm. In this case, linear regression assumes that there exists a linear relationship between the response variable and the explanatory variables.”

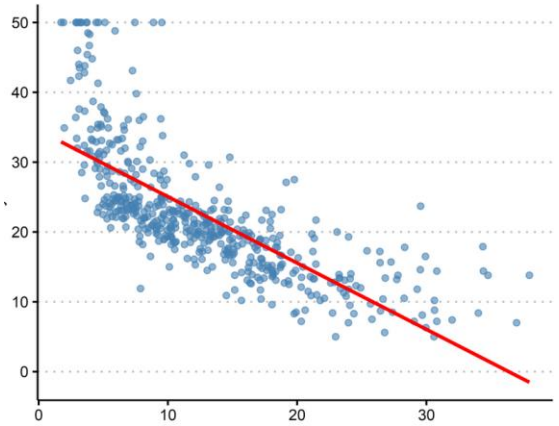
Source: “Linear Regression in R Tutorial”. July, 2018, accessed on September 29, 2022. https://www.datacamp.com/tutorial/linear-regression-R

Linear Regression

Linear relationships mean that “you can fit a line between the two (or more variables)”

- Take the example of the age and height we discussed before.





Source: “Linear Regression in R Tutorial”. July, 2018, accessed on September 29, 2022. https://www.datacamp.com/tutorial/linear-regression-R

Linear Regression

$$\text{Height} = a + b \times \text{Age}$$

Equation of a straight line: $Y = a + bx$

- “In this case, “a” and “b” are called the intercept and the slope respectively.”
- “With the same example, “a” or the intercept, is the value from where you start measuring. Newborn babies with zero months are not zero centimeters necessarily; this is the function of the intercept.”
- “The slope or “b” measures the change of height with respect to the age in months. In general, for every month older the child is, his or her height will increase with “b”.

Source: “Linear Regression in R Tutorial”. July, 2018, accessed on September 29, 2022. https://www.datacamp.com/tutorial/linear-regression-R

Linear Regression in R

- “A linear regression can be calculated in R with the command “lm”.”

lm([target variable] ~ [predictor variables], data = [data source])

```
Call:
lm(formula = height ~ age, data = ageandheight)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27238 -0.24248 -0.02762  0.16014  0.47238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.9283     0.5084  127.71  < 2e-16 ***
age           0.6350     0.0214   29.66 4.43e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.256 on 10 degrees of freedom
Multiple R-squared:  0.9888,    Adjusted R-squared:  0.9876
F-statistic:   880 on 1 and 10 DF,  p-value: 4.428e-11
```

Source: “Linear Regression in R Tutorial”. July, 2018, accessed on September 29, 2022. https://www.datacamp.com/tutorial/linear-regression-R

Linear Regression in R

Coefficients:

- “values of the intercept (“a” value)
- the slope (“b” value) for the age.
- These “a” and “b” values plot a line between all the points of the data.

  Height = a + b x Age

- There is a child that is 20.5 months old, a is 64.92 and b is 0.635, the model predicts (on average) that its height in centimeters is around 64.92 + (0.635 * 20.5) = 77.93 cm.”

```
Call:
lm(formula = height ~ age, data = ageandheight)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27238 -0.24248 -0.02762  0.16014  0.47238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.9283     0.5084  127.71  < 2e-16 ***
age           0.6350     0.0214   29.66 4.43e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.256 on 10 degrees of freedom
Multiple R-squared:  0.9888,    Adjusted R-squared:  0.9876
F-statistic:   880 on 1 and 10 DF,  p-value: 4.428e-11
```

Source: “Linear Regression in R Tutorial”. July, 2018, accessed on September 29, 2022. https://www.datacamp.com/tutorial/linear-regression-R

Linear Regression in R

- "When there is "two or more predictors to create the linear regression, it's called multiple linear regression."
- $Height = a + Age \times b_1 + (Number\ of\ Siblings) \times b_2$
- When comparing children with the same number of siblings, the average predicted height increases in 0.63 cm for every month the child age.
- The same way, when comparing children with the same age, the height decreases (because the coefficient is negative) in -0.01 cm for each increase in the number of siblings."

```
Call:
lm(formula = height ~ age + no_siblings, data = ageandheight)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28029 -0.22490 -0.02219  0.14418  0.48350

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.95872    0.55752 116.515 1.28e-15 ***
age          0.63516    0.02254  28.180 4.34e-10 ***
no_siblings -0.01137    0.05893  -0.193    0.851
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2693 on 9 degrees of freedom
Multiple R-squared:  0.9888,    Adjusted R-squared:  0.9863
F-statistic: 397.7 on 2 and 9 DF,  p-value: 1.658e-09
```

Source: "Linear Regression in R Tutorial". July, 2018, accessed on September 29, 2022. https://www.datacamp.com/tutorial/linear-regression-R

Linear Regression in R

p-values:

- “The blue rectangle indicates the *p*-values for the coefficients age and number of siblings. In simple terms, a p-value indicates whether or not you can reject or accept a hypothesis.
- Hypothesis: whether variables (age and no of siblings) are impacting the height of the children.

```
Call:
lm(formula = height ~ age + no_siblings, data = ageandheight)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28029 -0.22490 -0.02219  0.14418  0.48350

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.95872    0.55752 116.515 1.28e-15 ***
age          0.63516    0.02254  28.180 4.34e-10 ***
no_siblings -0.01137    0.05893  -0.193    0.851
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2693 on 9 degrees of freedom
Multiple R-squared:  0.9888,    Adjusted R-squared:  0.9863
F-statistic: 397.7 on 2 and 9 DF,  p-value: 1.658e-09
```

- The p-value for age is 4.34*e-10 or 0.000000000434. A very small value means that age is probably an excellent addition to your model.
- The p-value for the number of siblings is 0.85. In other words, there’s 85% chance that this predictor is not meaningful for the regression.”

Source: “Linear Regression in R Tutorial”. July, 2018, accessed on September 29, 2022. https://www.datacamp.com/tutorial/linear-regression-R

Linear Regression in R

p-values:

- “A standard way to test if the predictors are not meaningful is looking if the p-values smaller than 0.05.”
- In many cases, p-values smaller than 0.10 is also accepted.
- “This can visually interpreted by the significance stars at the end of the row against each X variable.
- The more the stars beside the variable’s p-Value, the more significant the variable.”

```
Call:
lm(formula = height ~ age + no_siblings, data = ageandheight)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28029 -0.22490 -0.02219  0.14418  0.48350

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.95872    0.55752 116.515 1.28e-15 ***
age          0.63516    0.02254  28.180 4.34e-10 ***
no_siblings -0.01137    0.05893  -0.193    0.851
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2693 on 9 degrees of freedom
Multiple R-squared:  0.9888,    Adjusted R-squared:  0.9863
F-statistic: 397.7 on 2 and 9 DF,  p-value: 1.658e-09
```

t-value: *t*-value needs to be greater than 1.96 to become *p*-value less than 0.05.

Source: “Linear Regression in R Tutorial”. July, 2018, accessed on September 29, 2022. https://www.datacamp.com/tutorial/linear-regression-R
Prabhakaran, S. “Complete Introduction to Linear Regression in R”. March 12, 2017. accessed on Oct 21, 2022. https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/

Linear Regression in R

Standard errors:

- “The standard deviation of an estimate is called the standard error.
- The standard error of the coefficient measures how precisely the model estimates the coefficient's unknown value.”

```
Call:
lm(formula = height ~ age + no_siblings, data = ageandheight)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28029 -0.22490 -0.02219  0.14418  0.48350

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.95872    0.55752 116.515 1.28e-15 ***
age          0.63516    0.02254  28.180 4.34e-10 ***
no_siblings -0.01137    0.05893  -0.193    0.851
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2693 on 9 degrees of freedom
Multiple R-squared:  0.9888,    Adjusted R-squared:  0.9863
F-statistic: 397.7 on 2 and 9 DF,  p-value: 1.658e-09
```

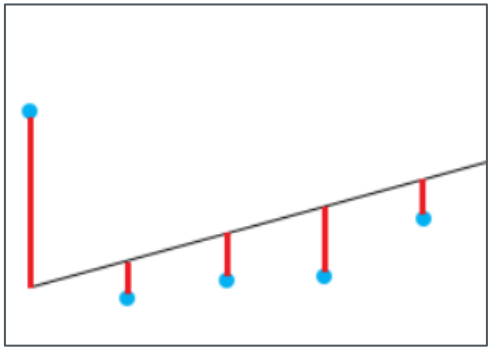
- We “use the standard error of the coefficient to measure the precision of the estimate of the coefficient. The smaller the standard error, the more precise the estimate”

Source: “What is the standard error of the coefficient?”. accessed on October 21, 2022. https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/regression/supporting-topics/regression-models/what-is-the-standard-error-of-the-coefficient/

Linear Regression in R

Residuals:

- Residuals are “the differences between the real values and the predicted values.
- The straight line in the image represents the predicted values.
- The red vertical line from the straight line to the observed data value is the residual.
- The idea in here is that the sum of the residuals is approximately zero or as low as possible. In real life, most cases will not follow a perfectly straight line, so residuals are expected.”

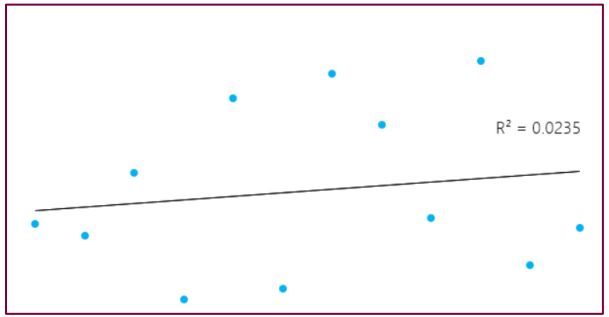


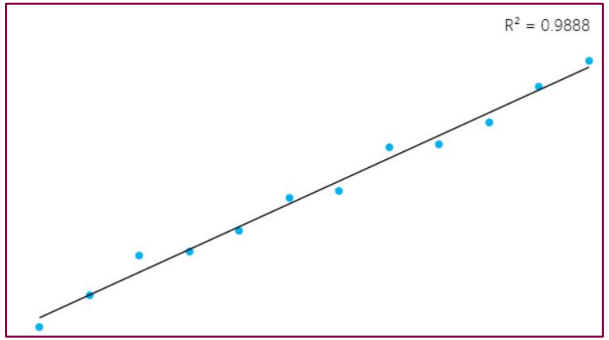
Source: “Linear Regression in R Tutorial”. July, 2018, accessed on September 29, 2022. https://www.datacamp.com/tutorial/linear-regression-R

Linear Regression in R

Model fit: *coefficient of determination or R^2*

- “For models that fit the data well, R^2 is near 1. Models that poorly fit the data have R^2 near 0.”
- “The first one has an R^2 of 0.02; this means that the model explains only 2% of the data variability.”
- “The second one has an R^2 of 0.99, and the model can explain 99% of the total variability.”
- “In real life, events don’t fit in a perfectly straight line all the time. For example, you can have in your data taller or smaller children with the same age. In some fields, an R^2 of 0.5 is considered good.”

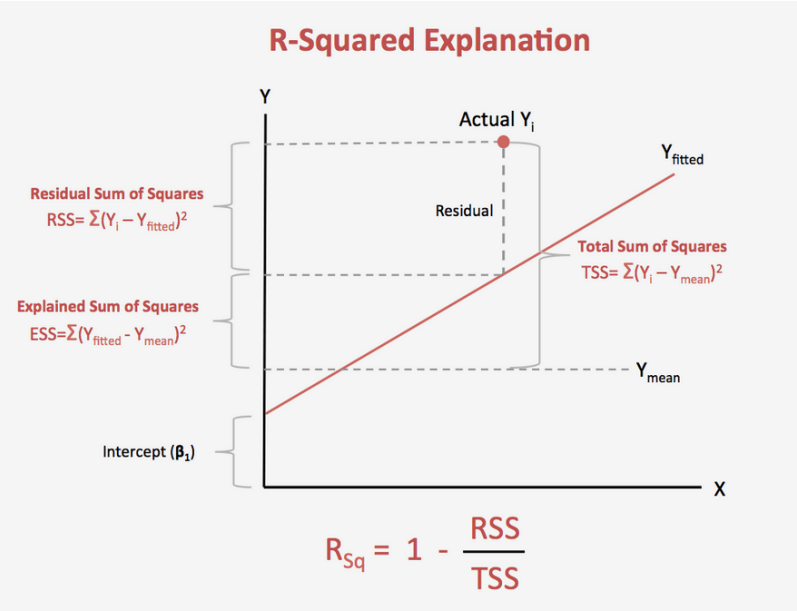




Source: “Linear Regression in R Tutorial”. July, 2018, accessed on September 29, 2022. https://www.datacamp.com/tutorial/linear-regression-R

Linear Regression in R

Model fit: *coefficient of determination or R^2*



Source: Prabhakaran, S. “Complete Introduction to Linear Regression in R”. March 12, 2017. accessed on Oct 21, 2022. https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/

Linear Regression in R

Model fit: *coefficient of determination or R^2*

- "In the blue rectangle, notice that there's two different R^2, one multiple and one adjusted.
- One problem with multiple R^2 is that it cannot decrease as you add more independent variables to your model, it will continue increasing as you make the model more complex, even if these variables don't add anything to your predictions (like the example of the number of siblings).
- For this reason, the adjusted R^2 is probably better to look at if you are adding more than one variable to the model, since it only increases if it reduces the overall error of the predictions."

```
Call:
lm(formula = height ~ age + no_siblings, data = ageandheight)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28029 -0.22490 -0.02219  0.14418  0.48350

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.95872    0.55752 116.515 1.28e-15 ***
age          0.63516    0.02254  28.180 4.34e-10 ***
no_siblings -0.01137    0.05893  -0.193    0.851
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2693 on 9 degrees of freedom
Multiple R-squared:  0.9888,    Adjusted R-squared:  0.9863
F-statistic: 397.7 on 2 and 9 DF,  p-value: 1.658e-09
```

Source: "Linear Regression in R Tutorial". July, 2018, accessed on September 29, 2022. https://www.datacamp.com/tutorial/linear-regression-R

Linear Regression in R

Checking the prediction accuracy of the model

Correlation between actual and predicted values:

- “A simple correlation between the actuals and predicted values can be used as a form of accuracy measure.
- A higher correlation accuracy implies that the actuals and predicted values have similar directional movement, i.e. when the actuals values increase the predicted values also increase and vice-versa.”

Source: Prabhakaran, S. “Complete Introduction to Linear Regression in R”. March 12, 2017. accessed on Oct 21, 2022. https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/

Linear Regression in R

Checking the prediction accuracy of the model

actuals	predicteds
116	172.9526
285	179.3271
48	162.8720
44	174.1385
98	183.4780
124	185.7016
238	167.1711
129	183.7744
166	181.6990
136	171.0254
139	170.5807
134	170.5807
285	173.2491
86	174.5833
154	181.1060

Min/ Max???

Min-Max accuracy:

- “Min-Max tells you how far the model's prediction is off. For a perfect model, this measure is 1.0. The lower the measure, the worse the model, based on out-of-sample performance.”

```
Min_Max Accuracy => mean(min(actual, predicted)/max(actual, predicted))
```

Source: “Meaning of Min-Max Accuracy of a regression model.” accessed on Oct 21, 2022. https://stats.stackexchange.com/questions/287143/meaning-of-min-max-accuracy-of-a-regression-model

Linear Regression in R

Checking the prediction accuracy of the model

Mean Absolute Percentage Error (MAPE:

- "The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics."
- "It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each case minus actual values divided by actual values."

$$\text{MAPE} = (1 / \text{sample size}) \times \sum[(|\text{actual} - \text{forecast}|) / |\text{actual}|] \times 100$$

Source: "What Is MAPE? (Plus How To Calculate MAPE in 3 Steps)" accessed on Oct 21, 2022. https://www.indeed.com/career-advice/career-development/what-is-mape

Workshop code!!!! :

Materials are available at: shorturl.at/aMT34

https://drive.google.com/drive/folders/1F5gW-DNz1NMiqQ6JJwFl7lxwYRXl02U0

Thank you!

Questions: jamals16@mcmaster.ca