How can we perform linear regression with categorical variables in R?

Authored by

stats writer

June 27, 2024

Linear regression is a statistical method used to examine the relationship between a dependent variable and one or more independent variables. In order to perform linear regression with categorical variables in R, we need to first convert the categorical variables into numerical values using techniques such as dummy coding or one-hot encoding. This allows us to incorporate the categorical variables into the regression model. Once the variables have been converted, we can use the lm() function in R to fit a linear regression model and obtain the corresponding coefficients and p-values. By including categorical variables in the regression model, we can better understand how they contribute to the relationship between the dependent and independent variables. This can help us make more accurate predictions and draw meaningful conclusions from our data.

Perform Linear Regression with Categorical Variables in R

Linear regression is a method we can use to quantify the relationship between one or more predictor variables and a .

Often you may want to fit a regression model using one or more as predictor variables.

This tutorial provides a step-by-step example of how to perform linear regression with categorical variables in R.

Example: Linear Regression with Categorical Variables in R

Suppose we have the following data frame in R that contains information on three variables for 12 different basketball players:

points scoredhours spent practicingtraining program used

#create data frame
df <- data.frame(points=c(7, 7, 9, 10, 13, 14, 12, 10, 16, 19, 22, 18),
hours=c(1, 2, 2, 3, 2, 6, 4, 3, 4, 5, 8, 6),
program=c(1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3))

#view data frame
df

points hours program
1 7 1 1
2 7 2 1
3 9 2 1
4 10 3 1
5 13 2 2
6 14 6 2
7 12 4 2
8 10 3 2
9 16 4 3
10 19 5 3
11 22 8 3

12 18 6 3

Suppose we would like to fit the following linear regression model:

points = β0 + β1hours + β2program

In this example, hours is a continuous variable but program is a categorical variable that can take on three possible categories: program 1, program 2, or program 3.

In order to fit this regression model and tell R that the variable "program" is a categorical variable, we must use as.factor() to convert it to a factor and then fit the model:

```
#convert 'program' to factor
df$program <- as.factor(df$program)

#fit linear regression model
fit <- lm(points ~ hours + program, data = df)

#view model summary
summary(fit)
```

Call:

lm(formula = points ~ hours + program, data = df)

Residuals:

Min 1Q Median 3Q Max

-1.5192 -1.0064 -0.3590 0.8269 2.4551

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.3013 0.9462 6.660 0.000159 *

hours 0.9744 0.3176 3.068 0.015401 *

program2 2.2949 1.1369 2.019 0.078234 .

**program3 6.8462 1.5499 4.417 0.002235 **

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Residual standard error: 1.403 on 8 degrees of freedom

Multiple R-squared: 0.9392, Adjusted R-squared: 0.9164

F-statistic: 41.21 on 3 and 8 DF, p-value: 3.276e-05

From the values in the Estimate column, we can write the fitted regression model:

points = 6.3013 + .9744(hours) + 2.2949(program 2) + 6.8462(program 3)

Here's how to interpret the coefficient values in the output:

hours: For each additional hour spent practicing, points scored increases by an average of 0.9744, assuming program is held constant.The p-value is .015, which indicates that hours spent practicing is a statistically significant predictor of points scored at level α = .05.program2: Players who used program 2 scored an average of 2.2949 more points than players who used program 1, assuming hours is held constant.The p-value is .078, which indicates that there is not a statistically significant difference in points scored by players who used program 2 compared to players who used program 1, at level α = .05.program3: Players who used program 3 scored an average of 2.2949 more points than players who used program 1, assuming hours is held constant.The p-value is .002, which indicates that there is a statistically significant difference in points scored by players who used program 3 compared to players who used program 1, at level α = .05.

Using the fitted regression model, we can predict the

number of points scored by a player based on their total hours spent practicing and the program they used.

```
#define new player
new <- data.frame(hours=c(5), program=as.factor(c(3)))

#use the fitted model to predict the points for the new player
predict(fit, newdata=new)

1
18.01923
```

The model predicts that this new player will score 18.01923 points.

We can confirm this is correct by plugging in the values for the new player into the fitted regression equation:

points = 6.3013 + .9744(hours) + 2.2949(program 2) + 6.8462(program 3)points = 6.3013 + .9744(5) + 2.2949(0) + 6.8462(1)points = 18.019

This matches the value we calculated using the predict() function in R.