



Statistical tools for high-throughput data analysis

Licence:



Search...



Home

Basics

Data

Visualize

Analyze

Products

Contribute

Support

About

Home / Articles / Machine Learning / Regression Analysis / Regression with Categorical Variables: Dummy Coding Essentials in R

Articles - Regression Analysis

Regression with Categorical Variables: Dummy Coding Essentials in R

[kassambara](#) | 11/03/2018 | 123997 | [Comments \(6\)](#) | [Regression Analysis](#)

This chapter describes how to compute **regression with categorical variables**.

Categorical variables (also known as *factor* or *qualitative variables*) are variables that classify observations into groups. They have a limited number of different values, called levels. For example the gender of individuals are a categorical variable that can take two levels: Male or Female.

Regression analysis requires numerical variables. So, when a researcher wishes to include a categorical variable in a regression model, supplementary steps are required to make the results interpretable.

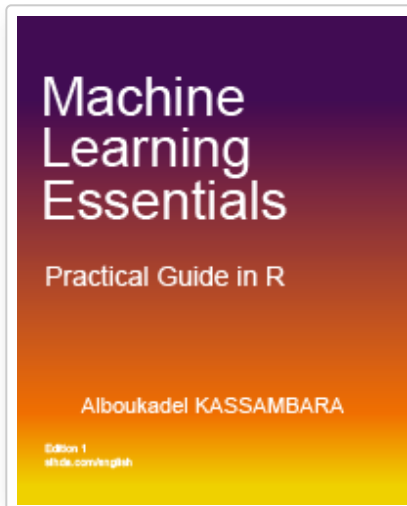
In these steps, the categorical variables are recoded into a set of separate binary variables. This recoding is called “dummy coding” and leads to the creation of a table called *contrast matrix*. This is done automatically by statistical software, such as R.

Here, you'll learn how to build and interpret a linear regression model with categorical predictor variables. We'll also provide practical examples in R.

Contents:

- [Loading Required R packages](#)
- [Example of data set](#)
- [Categorical variables with two levels](#)
- [Categorical variables with more than two levels](#)
- [Discussion](#)

The Book:



Machine Learning Essentials:
Practical Guide in R

Loading Required R packages

- **tidyverse** for easy data manipulation and visualization

```
library(tidyverse)
```

Example of data set

We'll use the **Salaries** data set [**car** package], which contains 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S.

The data were collected as part of the on-going effort of the college's administration to monitor salary differences between male and female faculty members.

```
# Load the data
data("Salaries", package = "car")
# Inspect the data
sample_n(Salaries, 3)
```

```
##      rank discipline yrs.since.phd yrs.service    sex salary
## 115 Prof           A           12           0 Female 105000
## 313 Prof           A           29           19  Male   94350
## 162 Prof           B           26           19  Male  176500
```

Categorical variables with two levels

Recall that, the regression equation, for predicting an outcome variable (y) on the basis of a predictor variable (x), can be simply written as $y = b_0 + b_1 \cdot x$. b_0 and b_1 are the regression beta coefficients, representing

the intercept and the slope, respectively.

Suppose that, we wish to investigate differences in salaries between males and females.

Based on the gender variable, we can create a new dummy variable that takes the value:

- 1 if a person is male
- 0 if a person is female

and use this variable as a predictor in the regression equation, leading to the following the model:

- $b_0 + b_1$ if person is male
- b_0 if person is female

The coefficients can be interpreted as follow:

1. b_0 is the average salary among females,
2. $b_0 + b_1$ is the average salary among males,
3. and b_1 is the average difference in salary between males and females.

For simple demonstration purpose, the following example models the salary difference between males and females by computing a simple linear regression model on the `Salaries` data set [`car` package]. R creates dummy variables automatically:

```
# Compute the model
model <- lm(salary ~ sex, data = Salaries)
summary(model)$coef
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   101002      4809    21.00 2.68e-66
## sexMale       14088       5065     2.78 5.67e-03
```

From the output above, the average salary for female is estimated to be 101002, whereas males are estimated a total of $101002 + 14088 = 115090$. The p-value for the dummy variable `sexMale` is very significant, suggesting that there is a statistical evidence of a difference in average salary between the genders.

The `contrasts()` function returns the coding that R have used to create the dummy variables:

```
contrasts(Salaries$sex)
```

```
##           Male
## Female      0
## Male        1
```

R has created a `sexMale` dummy variable that takes on a value of 1 if the sex is Male, and 0 otherwise. The decision to code males as 1 and females as 0 (baseline) is arbitrary, and has no effect on the regression computation, but does alter the interpretation of the coefficients.

You can use the function `relevel()` to set the baseline category to males as follow:

```
Salaries <- Salaries %>%
  mutate(sex = relevel(sex, ref = "Male"))
```

The output of the regression fit becomes:

```
model <- lm(salary ~ sex, data = Salaries)
summary(model)$coef
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   115090      1587    72.50 2.46e-230
## sexFemale     -14088      5065    -2.78 5.67e-03
```

The fact that the coefficient for `sexFemale` in the regression output is negative indicates that being a Female is associated with decrease in salary (relative to Males).

Now the estimates for `b0` and `b1` are 115090 and -14088, respectively, leading once again to a prediction of average salary of 115090 for males and a prediction of $115090 - 14088 = 101002$ for females.

Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable -1 (male) / 1 (female) . This results in the model:

- $b_0 - b_1$ if person is male
- $b_0 + b_1$ if person is female

So, if the categorical variable is coded as -1 and 1, then if the regression coefficient is positive, it is subtracted from the group coded as -1 and added to the group coded as 1. If the regression coefficient is negative, then addition and subtraction is reversed.

Categorical variables with more than two levels

Generally, a categorical variable with n levels will be transformed into $n-1$ variables each with two levels. These $n-1$ new variables contain the same information than the single variable. This recoding creates a table called **contrast matrix**.

For example `rank` in the `Salaries` data has three levels: "AsstProf", "AssocProf" and "Prof". This variable could be dummy coded into two variables, one called AssocProf and one Prof:

- If rank = AssocProf, then the column AssocProf would be coded with a 1 and Prof with a 0.
- If rank = Prof, then the column AssocProf would be coded with a 0 and Prof would be coded with a 1.
- If rank = AsstProf, then both columns "AssocProf" and "Prof" would be coded with a 0.

This dummy coding is automatically performed by R. For demonstration purpose, you can use the function `model.matrix()` to create a contrast matrix for a factor variable:

```
res <- model.matrix(~rank, data = Salaries)
head(res[, -1])
```

```
## rankAssocProf rankProf
## 1           0         1
## 2           0         1
## 3           0         0
## 4           0         1
## 5           0         1
## 6           1         0
```

When building linear model, there are different ways to encode categorical variables, known as contrast coding systems. The default option in R is to use the first level of the factor as a reference and interpret the remaining levels relative to this level.

Note that, ANOVA (analyse of variance) is just a special case of linear model where the predictors are categorical variables. And, because R understands the fact that ANOVA and regression are both examples of linear models, it lets you extract the classic ANOVA table from your regression model using the R base `anova()` function or the `Anova()` function [in `car` package]. We generally recommend the `Anova()` function because it automatically takes care of unbalanced designs.

The results of predicting salary from using a multiple regression procedure are presented below.

```
library(car)
model2 <- lm(salary ~ yrs.service + rank + discipline + sex,
             data = Salaries)
Anova(model2)
```

```
## Anova Table (Type II tests)
##
## Response: salary
##              Sum Sq  Df F value  Pr(>F)
## yrs.service 3.24e+08   1    0.63    0.43
## rank        1.03e+11   2  100.26 < 2e-16 ***
## discipline  1.74e+10   1   33.86 1.2e-08 ***
## sex         7.77e+08   1    1.51    0.22
## Residuals  2.01e+11 391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Taking other variables (yrs.service, rank and discipline) into account, it can be seen that the categorical variable sex is no longer significantly associated with the variation in salary between individuals. Significant variables are rank and discipline.

If you want to interpret the contrasts of the categorical variable, type this:

```
summary(model2)
```

```
##
## Call:
## lm(formula = salary ~ yrs.service + rank + discipline + sex,
##     data = Salaries)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -64202 -14255  -1533   10571   99163
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    73122.9     3245.3   22.53 < 2e-16 ***
## yrs.service     -88.8       111.6   -0.80  0.42696
## rankAssocProf  14560.4     4098.3    3.55  0.00043 ***
## rankProf       49159.6     3834.5   12.82 < 2e-16 ***
## disciplineB    13473.4     2315.5    5.82  1.2e-08 ***
## sexFemale      -4771.2     3878.0   -1.23  0.21931
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22700 on 391 degrees of freedom
## Multiple R-squared:  0.448, Adjusted R-squared:  0.441
## F-statistic: 63.4 on 5 and 391 DF, p-value: <2e-16
```

For example, it can be seen that being from discipline B (applied departments) is significantly associated with an average increase of 13473.38 in salary compared to discipline A (theoretical departments).

Discussion

In this chapter we described how categorical variables are included in linear regression model. As regression requires numerical inputs, categorical variables need to be recoded into a set of binary variables.

We provide practical examples for the situations where you have categorical variables containing two or more levels.

Note that, for categorical variables with a large number of levels it might be useful to group together some of the levels.

Some categorical variables have levels that are ordered. They can be converted to numerical values and used as is. For example, if the professor grades ("AsstProf", "AssocProf" and "Prof") have a special meaning, you can convert them into numerical values, ordered from low to high, corresponding to higher-grade professors.

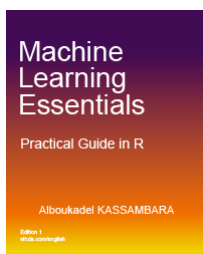
★ ★ ★ ★ ★ 5 Notes



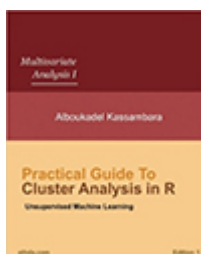
Enjoyed this article? Give us 5 stars ★ ★ ★ ★ ★ (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

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

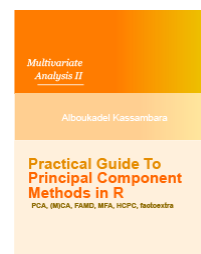
Recommended for You!



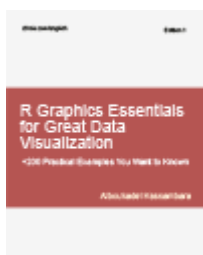
Machine Learning Essentials:
Practical Guide in R



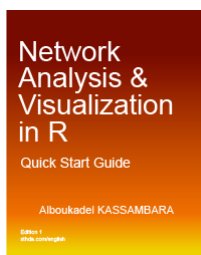
Practical Guide to Cluster
Analysis in R



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



You are not authorized to post a comment



Meny 04/03/2020 at 09h05
Member

I'd say this is just an answer to the person who asked "what does the value of (Intercept) Estimate= 73122.92 mean?"

The bias or intercept, in linear regression, is a measure of the mean of the response when all predictors are 0. That is, if you have $y = a + bx_1 + cx_2$, a is the mean y when x_1 and x_2 are 0.

#851



joanmelda 11/28/2019 at 06h36
Member

It is understandable that one is more confident when their task is in the hands of the Professional Essay Writing Help than a novice; thus, one hires Essay Writing Assignment Help Writer who delivers the ideal Custom Essay Paper
<https://superiorwriters247.com/essay-writing-help-services/>

#836



Visitor 04/27/2019 at 12h20

Visitor

thanks

#764

**tomer mann** 05/12/2018 at 15h00

Member

makes a hard concept easy!

#463

**kassambara** 03/23/2018 at 08h05

Administrator

In this example, intercept is the average salary we expect when all predictor variables equal zero.

Read more: [Interpreting Regression](#)

Thank you for your feedback

#398



SFer 03/22/2018 at 20h24

Visitor

Another clear and great article.
Thank you!.

In the very last part of the Article,
when you enter:

```
> summary(model2)
```

what does the value of

(Intercept) Estimate= 73122.92 mean?.

How to interpret this value: 73122.92
in the context of your example?
...(in simple, practical terms for an End User!).

pls, see code below:

```
summary(model2)
```

```
##  
## Call:  
## lm(formula = salary ~ yrs.service + rank + discipline + sex,  
## data = Salaries)  
##  
## Residuals:  
## Min 1Q Median 3Q Max  
## -64202 -14255 -1533 10571 99163  
##  
## Coefficients:  
## Estimate Std. Error t value Pr(> |t|)  
## (Intercept) 73122.9 3245.3 22.53 < 2e-16 ***  
## yrs.service -88.8 111.6 -0.80 0.42696  
## rankAssocProf 14560.4 4098.3 3.55 0.00043 ***  
## rankProf 49159.6 3834.5 12.82 < 2e-16 ***  
## disciplineB 13473.4 2315.5 5.82 1.2e-08 ***  
## sexFemale -4771.2 3878.0 -1.23 0.21931  
## ---  
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 22700 on 391 degrees of freedom  
## Multiple R-squared: 0.448, Adjusted R-squared: 0.441  
## F-statistic: 63.4 on 5 and 391 DF, p-value: <2e-16
```

#396


Sign in

Login

Password

Auto connect



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

Welcome!

Want to Learn More on R Programming and Data Science?

Follow us [by Email](#)

[Subscribe](#)

by [FeedBurner](#)

 [factoextra](#)

 [survminer](#)

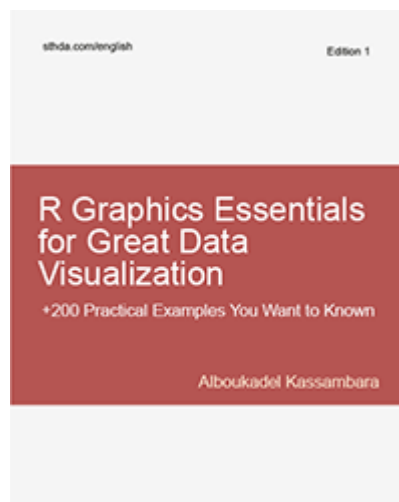
 [ggpubr](#)

 [ggcorrplot](#)

 [fastqcr](#)

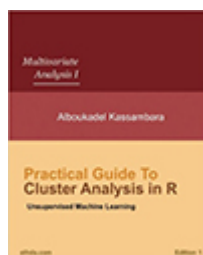
Our Books

3D Plots in R

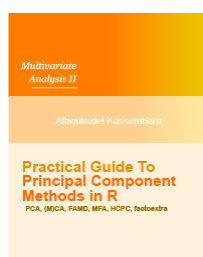


R Graphics Essentials for Great Data Visualization: 200 Practical Examples You Want to Know for Data Science

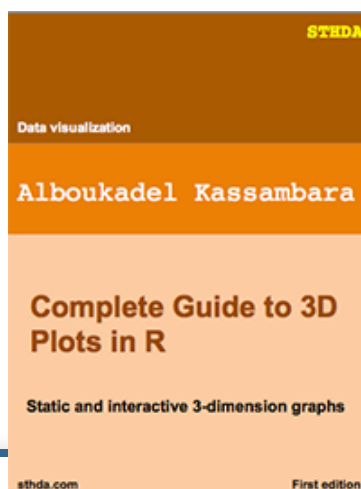
★ NEW!!



Practical Guide to Cluster Analysis in R



Practical Guide to Principal Component Methods in R



Newsletter

[Datanovia: Online Data Science Courses](#)[R-Bloggers](#)[Boosted by PHPBoost](#)