Overview

In the context of multiple regression analysis, we often think of predictor variables as being quantitative – e.g., annual salary, age of the study participants, number of years of education, sales revenue, and so on. In many ocassions, however, there exists the need of incorporating categorical predictor variables into a multiple regression model. For example, after controlling for differences attributable to, let's say, number of years of education, we may want to investigate whether there is a difference in salary between men and women. In this case, the variable gender is of interest and needs to be included in the regression model. Although the variable gender is not numerical, it can be given a numerical coding of some type so that it can be included in the model.

Binary (dichotomous) categorical variables can be represented in a regression model by using indicator or dummy variables. The standard approach is to code a binary variable with the values 0 and 1. For example, we can code a gender dummy variable with the value 0 for females and 1 for males; or use the value 0 for males and 1 for females. This lesson will demonstrate that these two different coding schemes yield the same analysis results. It is important to note that the coding values of 0 and 1 are easy to use and widely employed, but they are by no means the only way to code binary categorical variables.

Indicator variables can also be utilized in a regression equation to represent categorical variables with three or more categories or groups. As an example, suppose we have information that places a set of people in one of four race categories, namely Black, Hispanic, White, and Other. We can easily include this qualitative information along with quantitative variables in a regression model by using indicator variables.

In this lesson, we will utilize a dataset called *Salaries* – a dataset of the R *car* package – to demonstrate how to include categorical variables as predictors in a multiple regression model. We will also show that the particular dummy value we assign to each level of a categorical predictor makes no substantive difference in the results of the analysis. In each of the examples and activities presented in the lesson, we will assume that the response variable is a quantitative continuous variable.

We will start the lesson by describing the *Salaries* dataset. Then, we will present two basic examples on formulating and fitting regression models with one binary categorical predictor. In each of the examples, we will explain how to represent a binary categorical predictor through the use of a dummy variable. We will focus on the interpretation of the regression coefficients corresponding to a particular way of coding the dummy variable.

Right after the examples, the students will be asked to complete several activities. In them, the students will state and estimate the same regression models discussed in the examples, except that this time they will use a different way to code the binary categorical variable. The students will be guided to discover that the numerical values we choose to represent a categorical variable produce different regression coefficients, but has no effects in the ultimate results of the analysis.

At the very end of the lesson, the students will be given a set of homework exercises for them to complete. The exercises are intended to assess student knowledge and proficiency of the topics covered in the lesson.

The Salaries Dataset

As described in the R documentation, the *Salaries* dataset contains the 2008-09 nine-month academic salaries and other information for Assistant Professors, Associate Professors, and Professors at a particular college in the U.S. The data were collected as part of an on-going effort of the college administration to monitor salary differences between male and female faculty members.

Before running all the regression analyses for this unit, let's examine and describe the *Salaries* dataset. To get access to it, make sure you have loaded the *car* package into your R session. Let's start by using the dim( ) function to obtain the dimensions of the dataset.

```
dim(Salaries)
```

```
## [1] 397   6
```

The output indicates that *Salaries* has 397 observations and 6 variables. To get an idea of how the structure of the dataset looks like, let's apply the head( ) function to *Salaries*. This function, as coded below, displays the first twelve rows of the dataset.

```
knitr::kable(head(Salaries, n=12), caption = "First twelve rows of Salaries")
```

Table 1: First twelve rows of Salaries

rank	discipline	yrs.since.phd	yrs.service	sex	salary
Prof	B	19	18	Male	139750
Prof	B	20	16	Male	173200
AsstProf	B	4	3	Male	79750
Prof	B	45	39	Male	115000
Prof	B	40	41	Male	141500
AssocProf	B	6	6	Male	97000
Prof	B	30	23	Male	175000
Prof	B	45	45	Male	147765
Prof	B	21	20	Male	119250
Prof	B	18	18	Female	129000
AssocProf	B	12	8	Male	119800
AsstProf	B	7	2	Male	79800

We can obtain more information about the *Salaries* data by reading the *car* package documentation. This documentation describes each of the variables in the above table as follows:

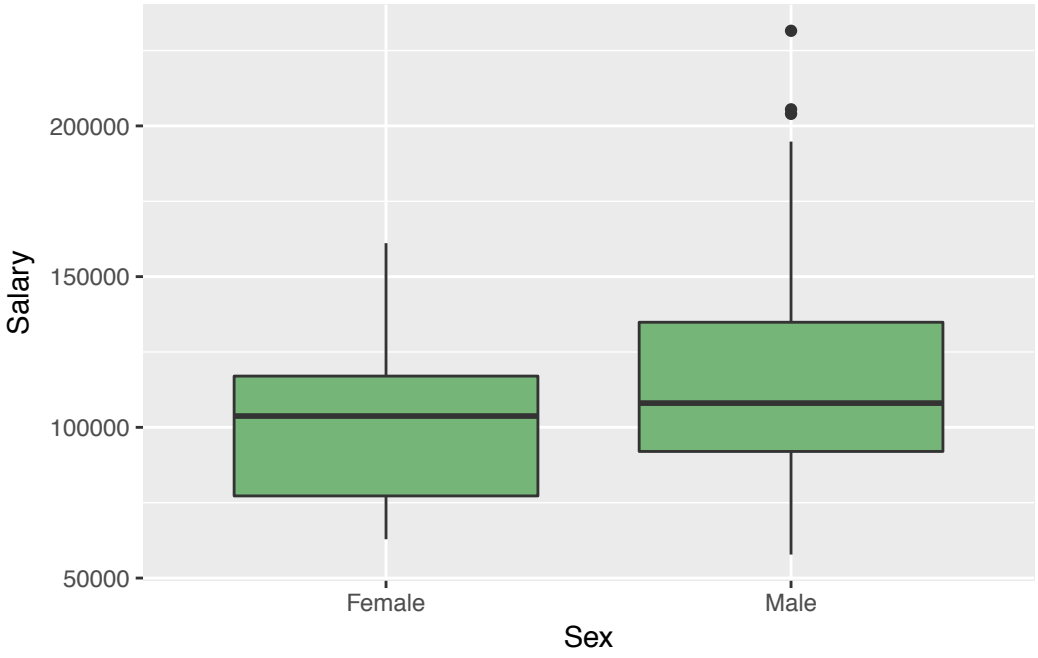
- *rank*: categorical variable with three levels – AssocProf (Associate Professor) , AsstProf (Assistant Professor), and Prof (Professor).
- *discipline*: categorical variable with two levels – A ("theoretical" departments) and B ("applied" departments).
- *yrs.since.phd*: years since PhD – a numerical variable.
- *yrs.service*: years of service – a numerical variable.

- *sex*: a categorical variable with two levels – Female and Male.
- *salary*: nine-month salary, in dollars – a numerical variable.

Example 1. Sex and Salary

To visualize the difference in salary between male and female instructors, let's make a call to the ggplot( ) function with the arguments Salaries and aes(x=sex, y=salary). This function call will display the pair of boxplots presented below.

```
ggplot(Salaries, aes(x=sex, y=salary)) +
    geom_boxplot(fill="#76b578") + labs(y="Salary", x="Sex")
```



From the boxplots, we can see that the median salary for male instructors is lightly higher than that of the female instructors. Since least-squares linear regressions are defined by the mean of the outcome variable, it would be useful to calculate and compare the mean salaries according to sex. The following line of code displays the mean salary both for male and for female.

```
means.sex <- tapply(salary, INDEX = sex, FUN=mean); means.sex
```

```
##   Female     Male
## 101002.4 115090.4
```

This output indicates that the mean salary for male is higher than the mean salary for female by $\$14,088$. Thus, the boxplots and the difference in means indicate, overall, that male instructors have a higher salary than female instructors – but is this difference in salary statistically significant? To answer this question, we are going to regress the variable *salary* on the binary categorical variable *sex* by using the lm( ) function. This function produces least squares estimates of the regression parameters. Before we perform this regression analysis, we will quantify the variable *sex* as 0 for

female and as 1 for male. Later on, you will be asked to run the analysis again and interpret the results, first, by making $sex = 1$ if the person is female and $sex = 0$ if the person is male (see Activity 1), and then, one more time, by making $sex = 1$ if the person is female and $sex = -1$ if the person is male (see Activity 2).

Regression of *salary* on *sex* – Coding Female as 0 and Male as 1:

The regression model we would like to fit can be stated as follows:

$$salary = \beta_0 + \beta_1 sex + \epsilon$$

The response function (i.e., the expected salary function) for this first-order regression model is:

$$E[salary] = \beta_0 + \beta_1 sex$$

For female instructors, $sex = 0$ and the response function becomes:

$$E[salary] = \beta_0$$

For male instructors, $sex = 1$ and the response function becomes:

$$E[salary] = \beta_0 + \beta_1$$

Therefore, the regression coefficients can be interpreted in the following way:

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

Next, let's use R to recode the variable *sex* as specified (0 for female and 1 for male).

```
# Recode "Female" as "0" and "Male" as "1":
sex <- as.character(sex)
sex[sex=="Female"] <- "0"
sex[sex=="Male"] <- "1"
sex <- as.factor(sex)

# Make sure that R uses the reference level of "0" (or, "Female") as the baseline:
sex <- relevel(sex, ref="0")
```

Finally, let's use the lm( ) function to fit the model and calculate the regression coefficients:

```
model1 <- lm(salary ~ sex, data=Salaries)
summary(model1)$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	101002.41	4809.386	21.001103	0.0000000
sexMale	14088.01	5064.579	2.781674	0.0056671

From this table, the average salary for female is estimated to be $b_0 = \$101,002.41$, while the average salary for male is approximately $b_0 + b_1 = \$101,002.41 + 14,088.01 = \$115,090.42$. The p-value corresponding to the dummy variable *sex* (*sexMale* in the table) is highly significant, which suggests that there exists statistical evidence of a difference in average salary between female and male faculty members.

Activity 1. Regression of *salary* on *sex* – Coding Female as 1 and Male as 0

In this activity, you are asked to fit and interpret the regression model in Example 1. This time, however, you will recode the variable *sex* such that $sex = 1$ if the person is female and $sex = 0$ if the person is male. Please accomplish the tasks and answer the questions presented below.

1. For the coding approach specified in this activity, write the expected salary function for female instructors and the expected salary function for male instructors.
2. How would you interpret the regression coefficients in this case? What is the meaning of b_0, b_1, and $b_0 + b_1$ in the context of this problem?
3. Use R to recode the variable *sex* as indicated in this activity. What would be the reference level or baseline in this case?
4. Once you recode the variable *sex* as requested, fit the model and calculate the regression coefficients by using the lm( ) function.
5. Obtain the values of b_0 and b_1 from the table displayed by the lm( ) function. What is the estimated average salary for female? What is the estimated average salary for male?
6. What is the p-value that tests the null hypethesis $H_0 : \beta_1 = 0$ against the altenative $H_0 : \beta_1 \neq 0$? Does the data provide sufficient evidence to conclude that there is a significant difference in salary between male and female instructors? Justify your answer.
7. Compare the analysis results of this activity to those of Example 1. Does it appear that they lead to the same conclusions in both cases? What changes and what remains the same? Do you think that the decision to code the variable *sex* as indicated in this activity changes the fundamental conclusions of the regression analysis? Explain.

Activity 2. Regression of *salary* on *sex* – Coding Female as 1 and Male as -1:

In this activity, carry out the same tasks and answer the same questions listed under Activity 1, but this time make $sex = 1$ if the person is female and $sex = -1$ if the person is male.

Example 2. Sex, Salary, and Years of Service

Let's examine the relationship between sex and salary further. We would expect salary to increase with years of service. Then, let's use our *Salaries* data to determine whether, after controlling for years of service, there is a difference in salary between male and female instructors. Specifically, let's consider a regression analysis to predict salary (*salary*) based on the number of years of service at the college (*yrs.service*) and gender (*sex*) of the instructor. Therefore, the first-order regression model is as follows:

$$salary = \beta_0 + \beta_1(yrs.service) + \beta_2 sex + \epsilon$$

To estimate this model, first, we will define our dummy variable by recoding "Female" as 0 and "Male" as 1. Later on (see Activity 3), you will be asked to repeat the entire regression analysis for this situation by assigning the 0s and 1s the other way around.

Regression of *salary* on *yrs.service* and *sex* – Coding Female as 0 and Male as 1

The response function (i.e., the expected salary function) for the above first-order regression model is:

$$E[salary] = \beta_0 + \beta_1(yrs.service) + \beta_2 sex$$

For female instructors, $sex = 0$ and the response function becomes:

$$E[salary] = \beta_0 + \beta_1(yrs.service)$$

For male instructors, $sex = 1$ and the response function becomes:

$$E[salary] = (\beta_0 + \beta_2) + \beta_1(yrs.service)$$

These two response functions represent parallel straight lines with different vertical intecepts. In this case, β_0 represents the average salary for female instructors and $\beta_0 + \beta_2$ represents the average salary for male instructors. Thus, if we speculate that male instructors might be getting a higher salary than female instructors, after controlling for years of service, we would expect the estimate of β_2 to be positive.

Next, let's use R to recode the variable *sex* as specified (0 for Female and 1 for Male).

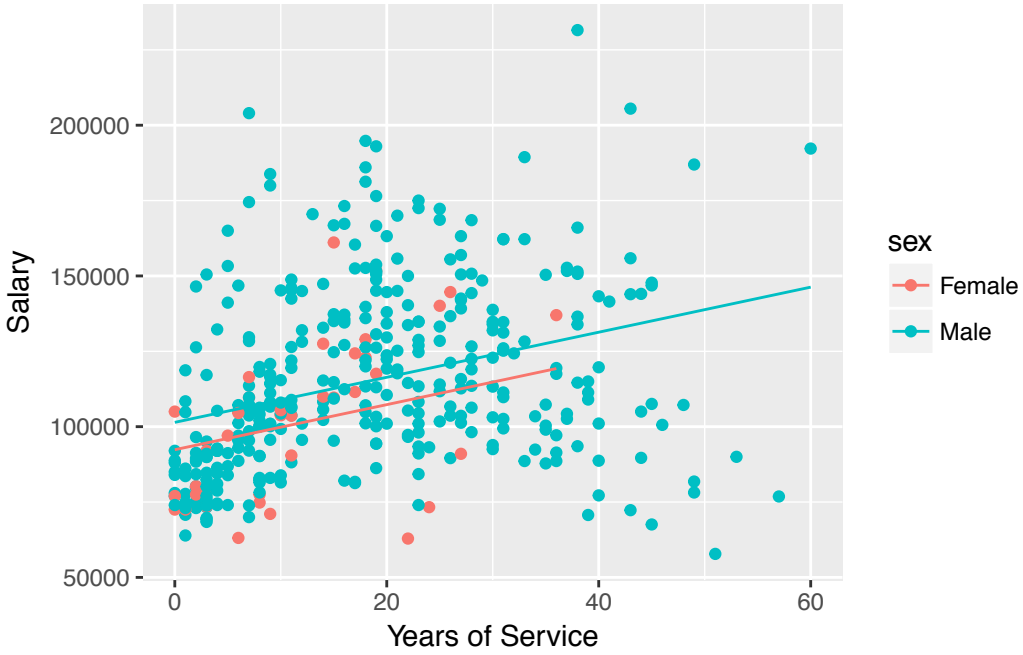
```
# Recode "Female" as "0" and "Male" as "1":
sex <- as.character(sex)
sex[sex=="Female"] <- "0"
sex[sex=="Male"] <- "1"
sex <- as.factor(sex)

# Make sure that R uses the reference level of "0" (or, "Male") as the baseline:
Salaries <- Salaries %>% mutate(sex = relevel(sex, ref="Female"))
```

A graph of the two response linear functions along with the scatterplot of salary versus years of service can be obtained with the following R code:

```
# Regression with same slope but different intercepts for each sex:
model3 <- lm(formula=salary ~ yrs.service + sex, data=Salaries)
Salaries = cbind(Salaries, pred = predict(model3))

ggplot(data=Salaries, mapping=aes(x=yrs.service, y=salary, color=sex)) +
    geom_point() +
    geom_line(mapping=aes(y=pred)) +
    labs(y="Salary", x="Years of Service")
```



This graph suggests that the salary for male is higher than the salary for female, on average. It also seems to indicate that β_2 is significant. By running the following regression analysis, which includes both years of service (*yrs.service*) and gender (*sex*) as predictors, we can both estimate the parameters of the model (β_0, β_1, and β_2) and test the null hypethesis $H_0 : \beta_2 = 0$ against the altenative $H_0 : \beta_2 \neq 0$.

```
model4 <- lm(salary ~ yrs.service + sex, data=Salaries)
summary(model4)$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	92356.9467	4740.188	19.483816	0.000000
yrs.service	747.6121	111.396	6.711301	0.000000
sexMale	9071.8000	4861.644	1.865994	0.062785

As you can see, men were paid more than women, after controlling for years of service. The *sex* coefficient (*sexMale* in the table) is positive with p-value = 0.062785. It appears that men were paid approximately $\$9,071.80$ more than women with the same number of years of service. After controlling for years of service, we can also read from the table that

- $b_0 = \$92,356.95$ is the estimated average salary for female instructors;
- $b_0 + b_2 = 92,356.95 + 9,071.80 = \$101,428.75$ is the estimated average salary for male instructors;
- $b_1 = \$747.61$ is the estimated change in average salary for an increase in one year of service;
- $b_2 = \$9,071.80$ is the estimated difference in average salary for male instructors as compared to female instructors, for a particular value of years of service.

Activity 3. Sex, Salary, and Years of Service – Making Female = 1, Male = 0

In Example 2, we recoded the variable *sex* by making $sex = 0$ if the person is female and $sex = 1$ if the person is male. In this activity, you will consider the same regression model as in Example 2 and recode the variable *sex* the other way around. That is, you will make $sex = 1$ if the person is female and $sex = 0$ if the person is male. Then, you will complete the tasks and answer the questions listed below.

1. For the recoding method specified in this activity, write the expected salary function for female and the expected salary function for male.
2. How would you interpret the regression coefficients in this case? What is the meaning of b_0, b_1, b_2, and $b_0 + b_2$ in the context of this problem?
3. Use R to recode the variable *sex* as indicated in this activity. What would be the reference level or baseline in this case?
4. Once you recode the variable *sex* as requested, use the ggplot( ) function to draw the salary function for male and the salary function for female as well as a scatterplot of salary versus years of service, all of them on the same Cartesian coordinate system. How does this graph look like in comparison to the one obtained in Example 2?
5. Fit the model and calculate the regression coefficients by using the lm( ) function.
6. Obtain the values of b_0, b_1, and b_2 from the table displayed by the lm( ) function. What is the estimated average salary for female instructors? What is the estimated average salary for male?
7. What is the estimated change in average salary per every year of service?
8. What is the estimated difference in average salary for male instructors as compared to female instructors for a fixed value of years of service?
9. What is the p-value that tests the null hypethesis $H_0 : \beta_2 = 0$ against the alternative $H_0 : \beta_2 \neq 0$? Does the data provide enough evidence to conclude that there is a significant difference in salary between male and female instructors? Justify your answer.
10. Compare the analysis results of this activity to those of Example 2. Does this activity help you understand that the manner in which we code a categorical variable in regression analysis is arbitrary? Please explain.