Dummy Variables for Dummies

Evgenia Olimpieva

02/25/2020

- So far we have only seen models with continuous independent variables.

- So far we have only seen models with continuous independent variables.
- Continuous (and discrete) variables are easiest to interpret

Continuous Independent Variable

- So far we have only seen models with continuous independent variables.
- Continuous (and discrete) variables are easiest to interpret
- Monthly Household Expense = $\beta_0 + \beta_1 *$Num of Kids$+ \ldots + \epsilon$

Continuous Independent Variable

- So far we have only seen models with continuous independent variables.
- Continuous (and discrete) variables are easiest to interpret
- Monthly Household Expense = $\beta_0 + \beta_1 *$Num of Kids$+ \ldots + \epsilon$
- We iterpret them in terms of the slope (change in run corresponding to a change in the rise)

- However, not all variables are interpreted in terms of the slope.

- However, not all variables are interpreted in terms of the slope.
- We interpret the coefficient of a dummy variable in the regression in terms of a difference between groups.

- However, not all variables are interpreted in terms of the slope.
- We interpret the coefficient of a dummy variable in the regression in terms of a difference between groups.
- Before we unpack that, let's review what dummy variables are!

- However, not all variables are interpreted in terms of the slope.
- We interpret the coefficient of a dummy variable in the regression in terms of a difference between groups.
- Before we unpack that, let's review what dummy variables are!
- Examples?

Indicator Variables

- Dummy variable = Indicator variable

- Dummy variable = Indicator variable
- Takes values 0 or 1.

- Dummy variable = Indicator variable
- Takes values 0 or 1.
- It indicates presence (1) or absence (0) of a certain quality.

Indicator variables

$$D = \begin{cases} 1 & \text{Property Present} \\ 0 & \text{Property Absent} \end{cases}$$

Indicator variables

$$D = \begin{cases} 1 & \text{Female} \\ 0 & \text{Male} \end{cases}$$

Creating an Indicator Variable

- Let's create an indicator variable!

Creating an Indicator Variable

- Let's create an indicator variable!
- We will use variable "English" and create a dummy variable "High Percent English Learners" or HiEL.

- Let's create an indicator variable!
- We will use variable "English" and create a dummy variable "High Percent English Learners" or HiEL.
- HiEL will take a value of 1 when there is a high percent of English learners in a school and 0 when it is low.

Formally:

$$\text{HiEL} = \begin{cases} 1 & \text{enlglish} \geq 10 \\ 0 & \text{enlglish} < 10 \end{cases}$$

Creating an Indicator Variable in R

We know how to do that in R!

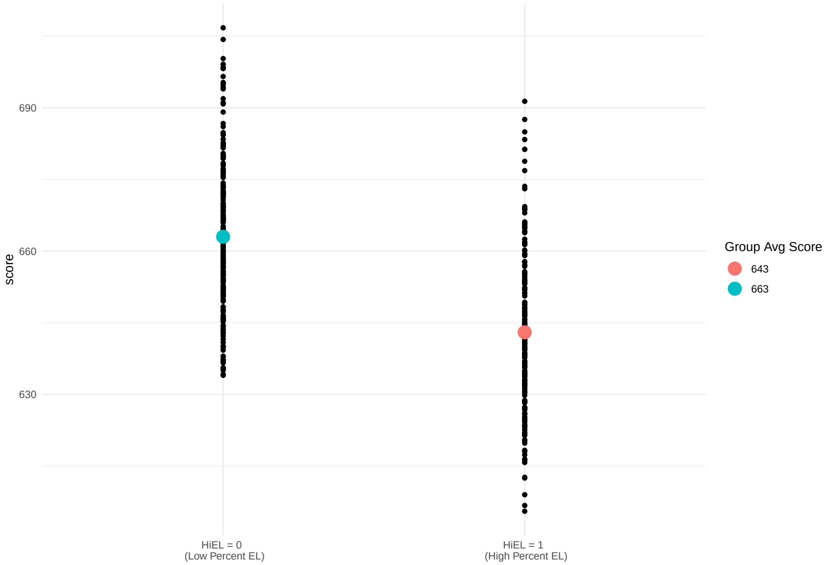
```
#create a dummy variable from variable `english`
CASchools$HiEL <- ifelse(CASchools$english >= 10, 1, 0)

#check your work
table(CASchools$HiEL)
```

```
##
##   0   1
## 228 192
```

Model with an Indicator Variable Only

$$\hat{\text{Score}} = \hat{\beta}_0 + \hat{\beta}_1 \overbrace{\text{HiEL}}^{\textit{Dummy}} + \epsilon$$

Model with an Indicator Variable Only



$$\text{Score} = \hat{\beta}_0 + \hat{\beta}_1 \text{HiEL} + \epsilon$$

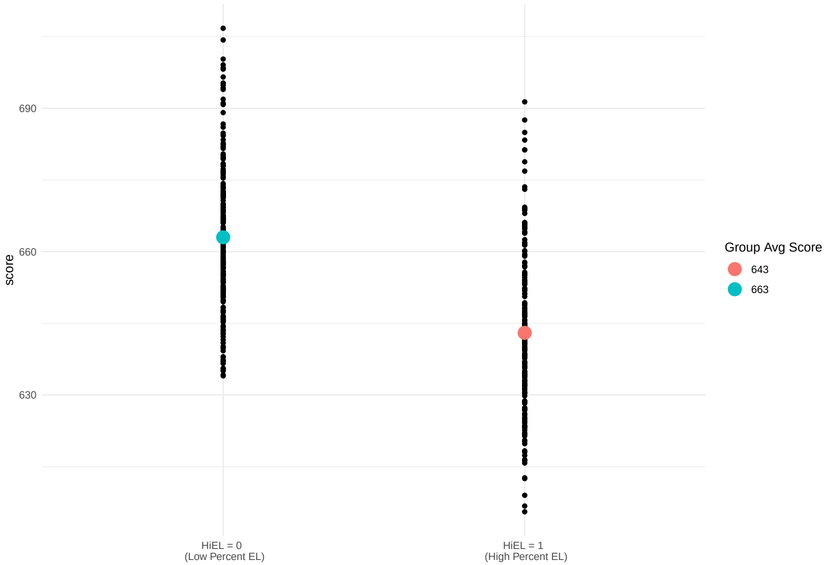
- For dummy-only model, it is not useful to think of $\hat{\beta}_1$ in terms of a slope
- $E(Y|HiEL = 0) = \hat{\beta}_0$
- $E(Y|HiEL = 1) = \hat{\beta}_0 + \hat{\beta}_1$
- Thus, β_1 is the difference in *group specific expectations*

Interpreting Models with an Indicator Variable Only

Table 1: Comparing Group Average Test Scores

	Dependent variable:
	score
HiEL	-20.400^{***}
	(1.580)
Constant	663.482^{***}
	(1.068)
Observations	420
R^2	0.285
Adjusted R^2	0.283
Residual Std. Error	16.129 (df = 418)
F Statistic	166.746^{***} (df = 1; 418)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Interpreting Models with an Indicator Variable Only



Indicator and Continuous Variable Model

- What if we added to our dummy-only regression a continous variable?

- What if we added to our dummy-only regression a continous variable?
- $Y = \beta_0 + \beta_X * \overbrace{X}^{Continuous} + \beta_D * \overbrace{D}^{Dummy} + \epsilon$

Indicator and Continuous Variable Model

- What if we added to our dummy-only regression a continous variable?
- $Y = \beta_0 + \beta_X * \overbrace{X}^{Continuous} + \beta_D * \overbrace{D}^{Dummy} + \epsilon$
- We will get something that is called a **parallel slopes model**, which you have seen on DataCamp.

Let´s add to our previous model a continous STR variable:

$$\hat{Score} = \hat{\beta}_0 + \hat{\beta}_1 * \overbrace{\text{HiEL}}^{\textit{Dummy}} + \hat{\beta}_2 * \overbrace{\text{STR}}^{\textit{Continuous}}$$

Model with and Indicator variable

Let's add to our previous model a continous STR variable:

$$\hat{Score} = \hat{\beta}_0 + \hat{\beta}_1 * \overbrace{\text{HiEL}}^{Dummy} + \hat{\beta}_2 * \overbrace{\text{STR}}^{Continuous}$$

What happens when HiEL= 0? What happens when HiEL=1?

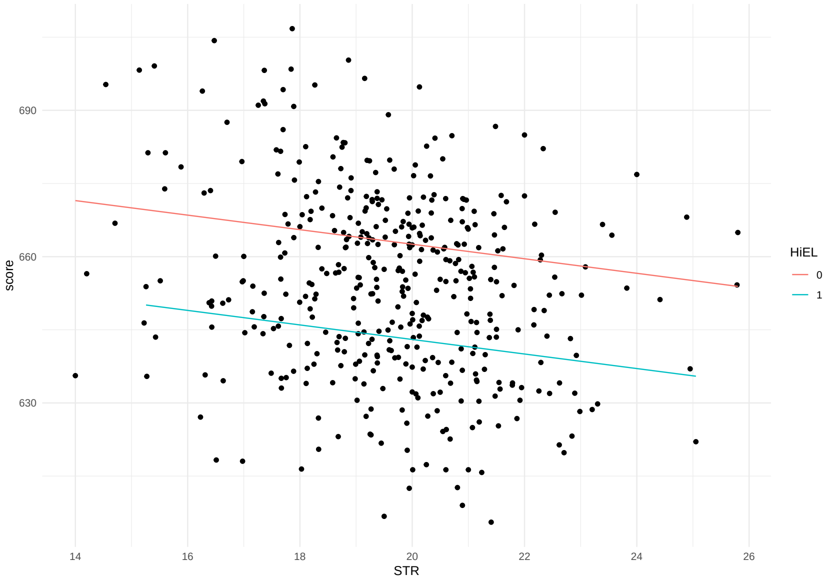
- By virtue of including the dummy and a continous variable, we created **two parallel slope models**, one for each group.
- One, for HiEL = 1, in which $\hat{\beta}_1$ gets added to the intercept β_0

$$\hat{Score} = (\hat{\beta}_0 + \hat{\beta}_1) + \hat{\beta}_2 * STR$$

- And another one for HiEL = 0, where $\hat{\beta}_1$ dissappears as it is multiplied by 0:

$$\hat{Score} = \hat{\beta}_0 + \hat{\beta}_2 * STR$$

Visualizing Indicator Variables in a Regression



Formally

$$Y = \beta_0 + \beta_D * \overbrace{D}^{indicator} + \beta_X * X + \epsilon$$

- If there is an indicator variable, our model essentially contains two models (one for each group): one for when D = 1, and another one for when D=0
- When $D = 0 \rightarrow Y = \beta_0 + \beta_X * X + \epsilon$. The coefficient for D is essentially added to the intercept (which is what shifts the line up and down).
- When $D = 1 \rightarrow Y = (\beta_0 + \beta_D) + \beta_X * X + \epsilon$

Table 2: Regression Results

	Dependent variable:
	score
HiEL	-19.533^{***}
	(1.576)
STR	-1.491^{***}
	(0.416)
Constant	692.361^{***}
	(8.121)
Observations	420
R^2	0.307
Adjusted R^2	0.303
Residual Std. Error	15.904 (df = 417)
F Statistic	92.170^{***} (df = 2; 417)
Note:	*p<0.1; **p<0.05; ***p<0.01

- Interpreting $\hat{\beta}_1 = -19.5$: On average, we expect the students in schools with High Percent English Learners to score 19.5 points **lower** on their test scores than students in school with Low Percent English Learners

Interpretation

- Interpreting $\hat{\beta}_1 = -19.5$: On average, we expect the students in schools with High Percent English Learners to score 19.5 points **lower** on their test scores than students in school with Low Percent English Learners
- Interpreting the intercept for HiEL = 0: "On average, when student to teacher ratio is zero, we expect students in schools with low number of english learners to have a test score around 692.4"

Interpretation

- Interpreting $\hat{\beta}_1 = -19.5$: On average, we expect the students in schools with High Percent English Learners to score 19.5 points **lower** on their test scores than students in school with Low Percent English Learners
- Interpreting the intercept for HiEL = 0: "On average, when student to teacher ratio is zero, we expect students in schools with low number of english learners to have a test score around 692.4"
- Interpreting the intercept for HiEL = 1: "On average, when student to teacher ratio is zero, we expect students in schools with high number of english learners to have a test score around 692.361-19.533= 672.8"

Interpretation

- Interpreting $\hat{\beta}_1 = -19.5$: On average, we expect the students in schools with High Percent English Learners to score 19.5 points **lower** on their test scores than students in school with Low Percent English Learners
- Interpreting the intercept for HiEL = 0: "On average, when student to teacher ratio is zero, we expect students in schools with low number of english learners to have a test score around 692.4"
- Interpreting the intercept for HiEL = 1: "On average, when student to teacher ratio is zero, we expect students in schools with high number of english learners to have a test score around 692.361-19.533= 672.8"
- The statistical significance of β_1 coefficient for `HiEL` in this case means that the difference between two groups is statistically significant.

Now, let's create a variable, which is the inverse of `HiEL` and which takes 1 when the school has less than 10 percent English learners and 0 otherwise.

$$\text{LowEL} = \begin{cases} 1 & \text{enlgish} < 10 \\ 0 & \text{english} \geq 10 \end{cases}$$

```r
#create a dummy variable from variable `english`
CASchools$LowEL <- ifelse(CASchools$english < 10, 1, 0)

#check your work
table(CASchools$LowEL)
```

```
##
##   0   1
## 192 228
```

Model with `LowEL`

$$\hat{Score} = \hat{\beta}_0 + \hat{\beta}_1 * \text{LowEL} + \hat{\beta}_2 * \text{STR}$$

We will run the same model as before, just replacing `HiEL` with `LowEL`

```
model_indicator_low <- lm(score ~ LowEL + STR,
                          data = CASchools)
```

Compare the two models

The two models below are **identical**, despite the fact that that 1) intercepts are different 2) the sign of coefficients for `HiEL` with `LowEL` is reversed. Why?

Table 3: Two Dummies Walked Into a Bar...

	Dependent variable:	
	score	
	(1)	(2)
HiEL	-19.533^{***}	
	(1.576)	
LowEL		19.533^{***}
		(1.576)
STR	-1.491^{***}	-1.491^{***}
	(0.416)	(0.416)
Constant	692.361^{***}	672.828^{***}
	(8.121)	(8.373)
Observations	420	420
R^2	0.307	0.307
Adjusted R^2	0.303	0.303
Residual Std. Error (df = 417)	15.904	15.904
F Statistic (df = 2; 417)	92.170^{***}	92.170^{***}
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01	

- We can never include both `HiEL` and `LowEL` because the two variables are perfectly collinear!

Beware of the Dummy Variable Trap!

- We can never include both `HiEL` and `LowEL` because the two variables are perfectly collinear!
- Lack of perfect collinearity is one of the key assumptions in linear models and we will go over it in more detail in our class on assumptions.

Beware of the Dummy Variable Trap!

- We can never include both `HiEL` and `LowEL` because the two variables are perfectly collinear!
- Lack of perfect collinearity is one of the key assumptions in linear models and we will go over it in more detail in our class on assumptions.
- However, an intuitive way to understand perfect collinearity is that it happens when one variable does not add any new information to another variable.

- All information contained in `LowEL` is already included in `HiEL` (they are the same variable, just coded differently). So, we do not need both of them in the model.

Beware of the Dummy Variable Trap!

- All information contained in `LowEL` is already included in `HiEL` (they are the same variable, just coded differently). So, we do not need both of them in the model.
- Adding both versions of the same dummy variables is dramatically referred to as a **dummy variable trap**

Beware of the Dummy Variable Trap!

Dummy variable trap is not so scarry. In fact, R simply won't let you add both of the variables into the model and will automatically drop one of them:

```
lm(score ~ LowEL + HiEL+ STR, data = CASchools)
```

```
##
## Call:
## lm(formula = score ~ LowEL + HiEL + STR, data = CASchool
##
## Coefficients:
## (Intercept)        LowEL         HiEL          STR
##     672.828       19.533           NA       -1.491
```

Thanks, R!