Categorical Variables in Regression Models

Jared S. Murray
The University of Texas at Austin
McCombs School of Business

Example: Estimating Wage Gaps

Imagine you are a trial lawyer and you are considering a suit against a company for salary discrimination. You've gathered the following data...

```
 Gender    Salary
1    Male   32.0
2  Female   39.1
3  Female   33.2
4  Female   30.6
5    Male   29.0
... ... ...
208 Female  30.0
```

Estimating Wage Gaps

You want to relate salary(Y) to gender(X)... how can we do that?

Gender is an example of a categorical variable. The variable gender separates our data into 2 groups or categories.

We want to understand the relationship between this categorical variable and salary.

Estimating Wage Gaps

Multiple regression will be useful here. First we recode the categorical variable into a dummy variable

```
 Gender    Salary  Male
1     Male  32.00   1
2   Female  39.10   0
3   Female  33.20   0
4   Female  30.60   0
5     Male  29.00   1
... ... ...
208  Female  30.00  0
```

Note: In R, categorical variables are known as **factors**. R will turn factor variables into dummies for you inside of `lm`

Estimating Wage Gaps

```
head(salary)
```

```
## # A tibble: 6 x 10
##   Employee EducLev JobGrade YrHired YrBorn Gender YrsPrior PCJob Salary   Exp
##      <int>   <int>    <int>   <int>  <int>  <chr>    <int> <chr>  <dbl> <dbl>
## 1        1       3        1      92     69   Male        1    No   32.0     4
## 2        2       1        1      81     57 Female        1    No   39.1    15
## 3        3       1        1      83     60 Female        0    No   33.2    13
## 4        4       2        1      87     55 Female        7    No   30.6     9
## 5        5       3        1      92     67   Male        0    No   29.0     4
## 6        6       3        1      92     71 Female        0    No   30.5     4
```

To ensure that Gender is treated as a categorical variable:

```
salary$Gender = factor(salary$Gender)
```

Estimating Wage Gaps

We could start by fitting the following model:

$$Salary_i = \beta_0 + \beta_1 Male_i + \epsilon_i$$

Estimating Wage Gaps

$$Salary_i = \beta_0 + \beta_1 Male_i + \epsilon_i$$

```
salaryfit = lm(Salary~Gender, data=salary)
coef(salaryfit)

## (Intercept)  GenderMale
##   37.209929    8.295513

confint(salaryfit)

##                 2.5 %   97.5 %
## (Intercept) 35.446314 38.97354
## GenderMale   5.211041 11.37998
```

How should we interpret these regression coefficients?

Estimating Wage Gaps

Plug in the two possible values for the dummy variable:

$$Salary_i = \begin{cases} 37.2 + \epsilon_i & \text{females} \\ 37.2 + 8.3 + \epsilon_i = 45.5 + \epsilon_i & \text{males} \end{cases}$$

```
mean(~Salary, data=subset(salary, Gender=="Female"))

## [1] 37.20993

mean(~Salary, data=subset(salary, Gender=="Male"))

## [1] 45.50544

print(45.50544 - 37.20993)

## [1] 8.29551
```

Estimating Wage Gaps

How can the defense attorney try to counteract the plaintiff's argument?

Perhaps the observed difference in salaries is due to confounding variables and NOT to gender discrimination...

Obviously, there are many other factors which we can legitimately use in determining salaries...

- education
- job classification
- experience
- ...

How can we use regression to incorporate additional information?

Estimating Wage Gaps

Let's add a measure of experience...

$$Salary_i = \beta_0 + \beta_1 Male_i + \beta_2 Exp_i + \epsilon_i$$

How do we interpret β_1 and β_2?

Estimating Wage Gaps

$$Salary_i = \beta_0 + \beta_1 Male_i + \beta_2 Exp_i + \epsilon_i$$

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.83075    1.08926  24.632  < 2e-16 ***
## GenderMale   8.01189    1.19309   6.715 1.81e-10 ***
## Exp          0.98115    0.08028  12.221  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.07 on 205 degrees of freedom
## Multiple R-squared:  0.491,Adjusted R-squared:  0.486
## F-statistic: 98.86 on 2 and 205 DF,  p-value: < 2.2e-16
```

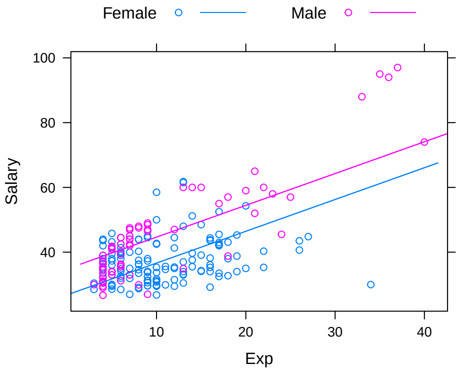
$$Salary_i = 27 + 8Male_i + 0.98Exp_i + \epsilon_i$$

How do we interpret these coefficients?

Estimating Wage Gaps

$$Salary_i = \begin{cases} 27 + 0.98 Exp_i + \epsilon_i & \text{females} \\ 35 + 0.98 Exp_i + \epsilon_i & \text{males} \end{cases}$$

```
plotModel(salaryfit_exp, Salary~Exp)
```



More than Two Categories

We can use dummy variables in situations in which there are more than two categories. Dummy variables are needed for each category except one, designated as the "baseline" or "reference" category.

The choice of reference category only effects the meaning of the coefficients in the model, not the overall fit (i.e. the fitted values, residual standard deviation, R^2, etc. remain the same)

Example: House Prices

We want to evaluate the difference in house prices in different neighborhoods.

	Nbhd	SqFt	Price
1	2	1.79	114.3
2	2	2.03	114.2
3	2	1.74	114.8
4	2	1.98	94.7
5	2	2.13	119.8
6	1	1.78	114.6
7	3	1.83	151.6
8	3	2.16	150.7
...	...	...	...

Example: House Prices

We could create dummy variables *dn*1, *dn*2 and *dn*3...

	Nbhd	SqFt	Price	dn1	dn2	dn3
1	2	1.79	114.3	0	1	0
2	2	2.03	114.2	0	1	0
3	2	1.74	114.8	0	1	0
4	2	1.98	94.7	0	1	0
5	2	2.13	119.8	0	1	0
6	1	1.78	114.6	1	0	0
7	3	1.83	151.6	0	0	1
8	3	2.16	150.7	0	0	1
...	...	...				

(Again, R will do this for you if you make Nbhd a factor)

Example: House Prices

$$Price_i = \beta_0 + \beta_1 dn2_i + \beta_2 dn3_i + \beta_3 Size_i + \epsilon_i$$

$$\begin{aligned} Price_i &= \beta_0 + \beta_3 Size + \epsilon_i && \text{(Nbhd 1)} \\ Price_i &= \beta_0 + \beta_1 + \beta_3 Size + \epsilon_i && \text{(Nbhd 2)} \\ Price_i &= \beta_0 + \beta_2 + \beta_3 Size + \epsilon_i && \text{(Nbhd 3)} \end{aligned}$$

Example: House Prices

$$Price = \beta_0 + \beta_1 dn2 + \beta_2 dn3 + \beta_3 Size + \epsilon$$

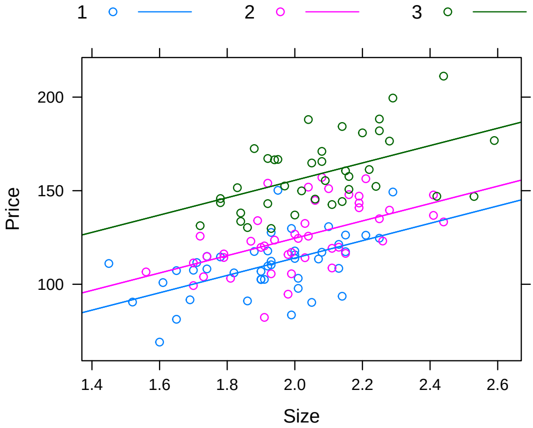
```
housing_fit = lm(Price~factor(Nbhd) + Size, data=housing)
coef(housing_fit)
```

```
##   (Intercept) factor(Nbhd)2 factor(Nbhd)3          Size
##         21.24         10.57         41.54         46.39
```

$$Price = 21.24 + 10.57dn2 + 41.54dn3 + 46.39Size + \epsilon$$

Example: House Prices

```
plotModel(housing_fit, Price~Size)
```



Example: House Prices

$$Price = \beta_0 + \beta_1 Size + \epsilon$$

```
lm(Price~Size, data=housing)

##
## Call:
## lm(formula = Price ~ Size, data = housing)
##
## Coefficients:
## (Intercept)         Size
##      -10.09        70.23
```

$$Price = -10.09 + 70.23 Size + \epsilon$$

Example: House Prices