Dummy Variables in Regression

R Tutorials for Applied Statistics

A **dummy variable** or **binary variable** is a variable that takes on a value of 0 or 1 as an indicator that the observation has some kind of characteristic. Common examples:

- Sex (female): FEMALE=1 if individual in the observation is female, equal to 0 otherwise
- Race (White): WHITE=1 if individual in the observation is white/Caucasian, equal to 0 otherwise
- Urban vs Rural: URBAN=1 if individual in the observation lives in an urban area, equal to 0 otherwise
- College graduate: COLGRAD=1 if individual in the observation has a four-year college degree, equal to 0 otherwise

It is common to use dummy variables as explanatory variables in regression models, if binary categorical variables are likely to influence the outcome variable.

1 Example: Factors Affecting Monthly Earnings

Let us examine a data set that explores the relationship between total monthly earnings (`MonthlyEarnings`) and a number of variables on an interval scale (i.e. numeric quantities) that may influence monthly earnings including including each person's IQ (`IQ`), a measure of knowledge of their job (`Knowledge`), years of education (`YearsEdu`), and years experience (`YearsExperience`), years at current job (`Tenure`).

The data set also includes dummy variables that may explain monthly earnings, including whether or not the person is black / African American (`Black`), whether or not the person lives in a Southern U.S. state (`South`), and whether or not the person lives in an urban area (`Urban`).

The code below downloads data on the above variables from 1980 for 663 individuals and assigns it to a dataframe called `df`.

```
load(url("http://murraylax.org/datasets/wage2.RData"))
```

The following call to `lm()` estimates a multiple regression predicting monthly earnings based on the eight explanatory variables given above, which includes three dummy variables. The next call to `summary()` displays some summary statistics for the estimated regression.

```r
lmwages <- lm(MonthlyEarnings
              ~ IQ + Knowledge + YearsEdu + YearsExperience + Tenure
              + Black + South + Urban,
              data = df)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure + Black + South + Urban, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -899.61 -228.64  -38.36  188.89 2138.04
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -475.254    143.374  -3.315 0.000968 ***
## IQ                 3.169      1.223   2.590 0.009816 **
## Knowledge          6.412      2.226   2.881 0.004097 **
## YearsEdu          45.563      8.664   5.259 1.97e-07 ***
## YearsExperience   13.356      3.969   3.365 0.000810 ***
## Tenure             3.711      2.954   1.256 0.209459
## Black           -107.905     55.539  -1.943 0.052460 .
## South            -37.840     31.072  -1.218 0.223732
## Urban            174.627     31.904   5.474 6.29e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 360.6 on 654 degrees of freedom
## Multiple R-squared:  0.2226, Adjusted R-squared:  0.2131
## F-statistic: 23.41 on 8 and 654 DF,  p-value: < 2.2e-16
```

The p-values in the right-most column reveal that all of the coefficients are statistically significantly different from zero at the 5% significance level. We have statistical evidence that all of these variables influence monthly earnings.

The coefficient on `Black` is equal to -107.91. This means that even after accounting for the effects of all the other explanatory variables in the model (includes educational attainment, experience, location, knowledge, and IQ), black / African American people earn on average $107.91 less per month than non-black people.

The coefficient on `South` is -37.84. Accounting for the impact of all the variables in the model, people that live in Southern United States earn on average $37.84 less per month than others.

The coefficient on `Urban` is 174.63. Accounting for the impact of all the variables in the model, people that live in urban areas earn $174.63 more per month, which probably reflects a higher cost of living.

We can compute confidence intervals for these effects with the following call to `confint()`

```
confint(lmwages, parm=c("Black", "South", "Urban"), level = 0.95)
```

```
##              2.5 %     97.5 %
## Black -216.96154   1.151043
## South  -98.85201  23.172580
## Urban  111.98018 237.273680
```

2 Dummy Interactions with Numeric Explanatory Variables

We found that black people have lower monthly earnings on average than non-black people. In our regression equation, this implies that the *intercept* is lower for black people than non-black people. We can also test whether a dummy variable affects the *slope* multiplying other variables.

For example, are there differences in the returns to education for black versus non-black people? To answer this, we include an *interaction effect* between `Black` and `YearsEdu`:

```
lmwages <- lm(MonthlyEarnings
              ~ IQ + Knowledge + YearsEdu + YearsExperience + Tenure
              + Black + South + Urban + Black*YearsEdu,
              data = df)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure + Black + South + Urban + Black * YearsEdu, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -899.84 -225.63  -43.73  185.46 2144.61
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -506.754    144.023  -3.519 0.000464 ***
## IQ                 3.118      1.221   2.553 0.010897 *
## Knowledge          6.755      2.228   3.032 0.002530 **
## YearsEdu          47.836      8.727   5.481 6.04e-08 ***
## YearsExperience   12.803      3.971   3.224 0.001328 **
## Tenure             3.600      2.948   1.221 0.222488
## Black            562.808    354.447   1.588 0.112805
## South            -36.621     31.015  -1.181 0.238128
## Urban            174.006     31.841   5.465 6.60e-08 ***
## YearsEdu:Black   -52.080     27.184  -1.916 0.055821 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 359.9 on 653 degrees of freedom
## Multiple R-squared:  0.227,  Adjusted R-squared:  0.2163
## F-statistic:  21.3 on 9 and 653 DF,  p-value: < 2.2e-16
```

We see here that when accounting for an interaction effect between race and education, the coefficient on the `Black` dummy variable becomes insignificant, but the coefficient on the interaction term is negative and significant at the 10% level. The coefficient on the interaction term equal to -52.08 means the slope on education is 52.08 less when $Black = 1$.

The coefficient on the interaction term is interpreted as the *additional* marginal effect of the numeric variable for the group associated with the dummy variable equal to 1. For this example:

- The marginal effect on monthly earnings for non-black people for an additional year of education is equal to $47.84 (i.e. when $Black = 0$).
- The marginal effect on monthly earnings for black people for an additional year of education is equal to $47.84 - $52.08 = -$4.24 (i.e. when $Black = 1$), which implies a near zero and possibly negative return to education on income for the black population (you would need to test the hypothesis for the linear combination to answer this).

3 Interacting Dummy Variables with Each Other

Let us interact two of the dummy variables to understand this interpretation and motivation. In the call to `lm()` below, we use our baseline model and interact `South` and `Urban`:

```
lmwages <- lm(MonthlyEarnings
              ~ IQ + Knowledge + YearsEdu + YearsExperience + Tenure
              + Black + South + Urban + South*Urban,
              data = df)
summary(lmwages)
```

```
## 
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience + 
##     Tenure + Black + South + Urban + South * Urban, data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -905.50 -222.90  -37.24  190.52 2128.95 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -538.952    148.641  -3.626 0.000310 ***
## IQ                  3.289      1.224   2.686 0.007405 ** 
## Knowledge           6.374      2.223   2.867 0.004278 ** 
## YearsEdu           46.735      8.685   5.381 1.03e-07 ***
## YearsExperience    14.011      3.985   3.516 0.000469 ***
## Tenure              3.675      2.950   1.246 0.213343    
## Black            -105.899     55.487  -1.909 0.056762 .  
## South              34.930     55.085   0.634 0.526234    
## Urban             213.078     39.922   5.337 1.30e-07 ***
## South:Urban      -104.974     65.652  -1.599 0.110315    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 360.2 on 653 degrees of freedom
## Multiple R-squared:  0.2256, Adjusted R-squared:  0.215 
## F-statistic: 21.14 on 9 and 653 DF,  p-value: < 2.2e-16
```

To interpret the meaning of the coefficient on `South`, `Urban`, and `South*Urban`, we will ignore (hold constant) all the terms in the regression equation that do not include one of these variables.

3.1 Difference between Urban and Rural Workers in the North/East/West

Workers in the North / East / and West U.S. have $South = 0$. Here $South = 0$, $(South \text{ x } Urban) = 0$, so neither the coefficient on the interaction nor the coefficient on `South` come into play.

The coefficient for b_{Urban} implies that *in the Non-Southern U.S.*, urban workers earn on average $213.08 more in monthly earnings than rural workers.

3.2 Difference between Urban and Rural Workers in the South

When focusing on workers in the South, $South = 1$ and the interaction term comes into play.

- Impact for urban workers in the south = $b_{South}(1) + b_{Urban}(1) + b_{Urban*South}(1)$
- Impact for rural workers in the south = $b_{South}(1) + b_{Urban}(0) + b_{Urban*South}(0)$
- Difference = $b_{Urban} + b_{Urban*South} = 213.08 - 104.97 =$ $108.11

In the Southern U.S. states, urban workers on average earn $108.11 more in monthly earnings than rural workers.

3.3 Difference between Southern and North/East/West Monthly Earnings for ***Urban*** Workers

- Impact for Southern urban workers = $b_{South}(1) + b_{Urban}(1) + b_{Urban*South}(1)$
- Impact for Non-Southern urban workers = $b_{South}(0) + b_{Urban}(1) + b_{Urban*South}(0)$
- Difference = $b_{South} + b_{Urban*South} = 34.93 - 104.97 =$ -$70.04

For urban workers, workers in the South earn $70.04 less in monthly earnings than workers in the North/East/West.

3.4 Difference between Southern and North/East/West Monthly Earnings for *Rural* Workers

Rural workers have $\text{Urban} = 0$ and so the interaction term $\text{Urban x South} = 0$, so we can ignore both of those coefficients. The coefficient for b_{South} implies that Southern rural workers earn on average 34.93 *more* per month than Non-Southern rural workers.

3.5 Three-Way Interactions and Higher!

What?! Things aren't complicated enough for you?! Do at your own peril!

I have seen people include higher order interaction effects like `South * Urban * Black * YearsEdu` in their regressions. It has never been obvious to me that they understood what their results meant.