Dummy Variables

ECO 6416

2022-10-04

Contents

Here are all the packages needed to get started.

```
library(readxl)



sessionInfo()
```

```
## R version 4.2.1 (2022-06-23 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19044)
## 
## Matrix products: default
## 
## locale:
## [1] LC_COLLATE=English_United States.utf8 
## [2] LC_CTYPE=English_United States.utf8   
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C                          
## [5] LC_TIME=English_United States.utf8    
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
## [1] readxl_1.4.1
## 
## loaded via a namespace (and not attached):
##  [1] digest_0.6.29    cellranger_1.1.0 magrittr_2.0.3   evaluate_0.15   
##  [5] rlang_1.0.3      stringi_1.7.8    cli_3.3.0        rstudioapi_0.14 
##  [9] rmarkdown_2.14   tools_4.2.1      stringr_1.4.0    xfun_0.31       
## [13] yaml_2.3.5       fastmap_1.1.0    compiler_4.2.1   htmltools_0.5.2 
## [17] knitr_1.39
```

1 Influences on SAT Scores

Suppose you are interested in the determinants of SAT scores.

```
sat <- read_xlsx("../Data/8SAT.xlsx")

sat_86AP <- sat[,-1]
```

I am temporarily dropping the `AP` column because I want to demonstrate something first.

With all the data, you decide to run a full model with all the independent variables. The `~.` is a nice trick. It includes all the remaining variables from the dataframe.

```
model_full <- lm(SAT ~., data = sat_86AP)

summary(model_full)
```

```
##
## Call:
## lm(formula = SAT ~ ., data = sat_86AP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -281.06  -63.05   -0.61   69.03  362.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   542.47     113.02   4.800 1.19e-05 ***
## APENG          80.32      45.64   1.760 0.083814 .
## APMATH         80.48      40.54   1.985 0.051930 .
## ESL            27.97      61.96   0.451 0.653456
## GEND          108.38      30.97   3.500 0.000911 ***
## GPA           130.09      37.34   3.484 0.000957 ***
## PREP          -40.50      36.12  -1.121 0.266842
## RACE          -79.07      65.34  -1.210 0.231223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 123.2 on 57 degrees of freedom
## Multiple R-squared:  0.6309, Adjusted R-squared:  0.5855
## F-statistic: 13.92 on 7 and 57 DF,  p-value: 2.321e-10
```

1.1 Interpreting Results

Since Gender and GPA are the only ones that tested significant at the 95% level, let's analyze the coefficients. If a person's GPA increases by 1, SAT scores are expected to increase by 130 points, give or take 74 points.

In this case, Gender equaling 1 means that a female took the test. When interpreting, we would say that females score about 108 points higher than males give or take 60 points.

1.2 Prediction

Let's predict a student's SAT score based on their characteristics. We could build a prediction dataframe and use that, but for simplicity, I am just going to grab a single row instead of building one.

```
newPrediction <- sat[8,]

predict(model_full, newdata= newPrediction, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 884.4303 820.5172 948.3435
```

1.3 Partial F-Test

Another question that might be asked is if including APMath and APEnglish reduced the SSE by a statistically significant amount.

We can run a reduced model, then compare the two using a partial F test.

1.3.1 Building Reduced Model

```
reduced_sat <- sat[,-(1:3)] # drops first 3 columns

model_reduced <- lm(SAT ~. , data = reduced_sat)

summary(model_reduced)
```

```
## 
## Call:
## lm(formula = SAT ~ ., data = reduced_sat)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -366.01  -78.25   17.55   72.56  357.50
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   365.40      99.02   3.690 0.000491 ***
## ESL            72.12      63.52   1.135 0.260824
## GEND          113.20      32.38   3.496 0.000902 ***
## GPA           206.64      26.99   7.657 2.08e-10 ***
## PREP          -50.82      37.81  -1.344 0.184026
## RACE          -98.15      67.28  -1.459 0.149912
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 130.1 on 59 degrees of freedom
## Multiple R-squared:  0.5737, Adjusted R-squared:  0.5376
## F-statistic: 15.88 on 5 and 59 DF,  p-value: 6.873e-10
```

1.3.2 Comparing Models

Now we can compare the reduction in both models using a partial F test

```
anova(model_reduced, model_full)
```

```
## Analysis of Variance Table
## 
## Model 1: SAT ~ ESL + GEND + GPA + PREP + RACE
## Model 2: SAT ~ APENG + APMATH + ESL + GEND + GPA + PREP + RACE
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     59 998969
## 2     57 865071  2    133899 4.4113 0.01655 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Neat! Including the two did reduce the SSE by a statistically significant amount, but independently, neither tested statistically significant.

We could have also seen this if we used the `AP` column instead of breaking apart what AP subject was taken.

```
sat_AP <- sat[,-(2:3)]

ap_model <- lm(SAT ~. , data = sat_AP)

summary(ap_model)
```

```
##
## Call:
## lm(formula = SAT ~ ., data = sat_AP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -255.78  -70.18  -16.40   69.25  315.45
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   473.13      96.17   4.920 7.51e-06 ***
## AP            144.37      41.83   3.451  0.00105 **
## ESL            32.61      59.47   0.548  0.58552
## GEND          126.16      29.98   4.208 9.05e-05 ***
## GPA           144.16      30.70   4.696 1.67e-05 ***
## PREP          -35.15      35.03  -1.003  0.31979
## RACE          -90.44      61.85  -1.462  0.14904
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 119.5 on 58 degrees of freedom
## Multiple R-squared:  0.6464, Adjusted R-squared:  0.6098
## F-statistic: 17.67 on 6 and 58 DF,  p-value: 1.63e-11
```

This means, that if you took an AP course, you are likely to increase your SAT scores by 144 points, give or take 84 points.

How is that possible? This simply means if you take an AP class, your scores are likely to increase, but there is no discernible difference between taking APMath, APEnglish, or both.