Middle School Regression Analysis

Andrea Tillotson

4/29/2022

Logistic Regression on More than Five Tested

Model (in logit terms): not easily interpretable

This first model regresses the binary variable of whether there are **more than five testers** or not on the categorical variable of what borough a school is in.

The baseline is now 0 (rather than the Bronx), so the values are relative to 0.

For the Bronx coefficient: A school being in the Bronx increases the **log-odds (logit)** that the school had more than five test takers by 1.37. A logit is the ratio of two probabilities and isn't very interpretable.

```
#running logit model
logit <- glm(MTF_testers ~ 0 + Borough, family = "binomial",
    data = MS_data)
logit %>% summary()
```

```
##
## Call:
## glm(formula = MTF_testers ~ 0 + Borough, family = "binomial",
##     data = MS_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9186   0.5879   0.6716   0.6726   0.7868
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## BoroughBronx           1.3710     0.1949   7.034 2.01e-12 ***
## BoroughBrooklyn        1.3743     0.1727   7.957 1.77e-15 ***
## BoroughManhattan       1.0141     0.1919   5.284 1.26e-07 ***
## BoroughQueens          1.6677     0.2438   6.841 7.88e-12 ***
## BoroughStatenIsland    1.1632     0.5123   2.270   0.0232 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 910.8  on 657  degrees of freedom
## Residual deviance: 667.9  on 652  degrees of freedom
## AIC: 677.9
##
## Number of Fisher Scoring iterations: 4
```

Odds ratios: not easily interpretable

Here, I get the coefficients from above in terms of the **odds ratio** rather than log-odds (formula = $e^{logit\ coefficient}$). Keep in mind that with odds ratios, a number less than 1 is a decrease in odds and a number greater than 1 is an increase in odds.

A school being in the Bronx increases the **odds** that the school had more than five test takers by about 293.94% (3.9394 − 1 or $e^{logit\ coefficient} - 1$). Note, this is still not a probability but a proportionate change in odds.

```
# getting odds ratios
exp(coef(logit))
```

```
##        BoroughBronx     BoroughBrooklyn    BoroughManhattan       BoroughQueens
##            3.939394            3.952381            2.756757            5.300000
## BoroughStatenIsland
##            3.200000
```

Predicted probabilities: most interpretable

Below, we finally arrive at predicted probabilities, which are probably the most interpretable way to report results for general audiences.

As we can see, the predicted probability that a school in the Bronx has more than five test takers is about 79.75%. A school in Brooklyn's predicted probability of having more than five test takers is 79.81%. For Manhattan, the predicted probability is 73.38%, for Queens the predicted probability is 84.13%, and for Staten Island the predicted probability is 76.19%.

```
newdata <- with(MS_data, data.frame(Borough = factor(c("Bronx", "Brooklyn", "Manhattan",
                                                       "Queens", "StatenIsland"))))
predProb <- as.data.frame(predict(logit, newdata, type = "response")) %>%
  magrittr::set_rownames(c("Bronx", "Brooklyn", "Manhattan", "Queens", "StatenIsland")) %>%
  transmute(Predicted_Probabilities = `predict(logit, newdata, type = "response")`)

predProb
```

```
##             Predicted_Probabilities
## Bronx                     0.7975460
## Brooklyn                  0.7980769
## Manhattan                 0.7338129
## Queens                    0.8412698
## StatenIsland              0.7619048
```

Same process for more than five offers

The code here is the same as for the first model so I'll just include the output.

Model (in logit terms): not easily interpretable

We can see that all the boroughs' coefficients are statistically significant for more than five offers. We also see that, even though a school being in the Bronx had **increased** its log odds of having more than five

testers, it significantly **decreases** its log odds of having more than five offers. I wonder if there have been community or local efforts in the Bronx to get more students tested? But if this hasn't translated in offers.

```
##
## Call:
## glm(formula = MTF_offers ~ 0 + Borough, family = "binomial",
##     data = MS_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9005  -0.6436  -0.5975  -0.3172   2.4553
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## BoroughBronx          -2.9640     0.3625  -8.176 2.94e-16 ***
## BoroughBrooklyn       -1.6327     0.1875  -8.707  < 2e-16 ***
## BoroughManhattan      -1.4693     0.2175  -6.755 1.43e-11 ***
## BoroughQueens         -0.8777     0.1956  -4.487 7.22e-06 ***
## BoroughStatenIsland   -0.6931     0.4629  -1.497    0.134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 910.80  on 657  degrees of freedom
## Residual deviance: 562.37  on 652  degrees of freedom
## AIC: 572.37
##
## Number of Fisher Scoring iterations: 5
```

Odds ratios: not easily interpretable

Here, we can see how significant a school being in the Bronx is for having more than five offers. A school being in the Bronx decreases the odds that the school had more than five offers takers by about 94.84% $(0.0516 - 1)$ or $e^{logit\ coefficient} - 1)$.

```
##        BoroughBronx     BoroughBrooklyn    BoroughManhattan       BoroughQueens
##           0.0516129           0.1954023           0.2300885           0.4157303
## BoroughStatenIsland
##           0.5000000
```

Predicted probabilities: most interpretable

The predicted probability that a school in _____ has more than five offers is about _____:

- Bronx, 4.91%
- Brooklyn, 16.35%
- Manhattan, 18.71%
- Queens, 29.37%
- Staten Island, 33.33%

```
##              Predicted_Probabilities
## Bronx                     0.04907975
## Brooklyn                  0.16346154
## Manhattan                 0.18705036
## Queens                    0.29365079
## StatenIsland              0.33333333
```

Linear regression for percent of pool tested against borough

The Bronx is the intercept here and all other values are compared to the Bronx. Positive values for other boroughs indicate that, for schools in that borough, the percentage of the pool tested is higher on average than the Bronx schools. Negative values indicate that, for schools in that borough, the percentage of the pool tested is lower than that of the Bronx.

A school being in Manhattan, for example, increases the percent of the pool tested by about 11.25 compared to the Bronx. Conversely, a school being in Staten Island decreases the average percent of the pool tested by about 2.55 compared to the Bronx.

Keep in mind that this regression output does not account for schools with fewer than five test takers (see the 137 observations deleted noted in the output).

```
lm(`Percent of pool tested` ~ Borough, data = MS_data) %>% summary()
```

```
##
## Call:
## lm(formula = `Percent of pool tested` ~ Borough, data = MS_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.840 -13.695  -4.363  10.713  65.005
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           26.695      1.622  16.456  < 2e-16 ***
## BoroughBrooklyn        5.168      2.166   2.386   0.0174 *
## BoroughManhattan      11.246      2.447   4.597 5.41e-06 ***
## BoroughQueens          5.862      2.421   2.422   0.0158 *
## BoroughStatenIsland   -2.551      4.900  -0.521   0.6029
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.5 on 515 degrees of freedom
##   (137 observations deleted due to missingness)
## Multiple R-squared:  0.0446, Adjusted R-squared:  0.03718
## F-statistic:  6.01 on 4 and 515 DF,  p-value: 9.873e-05
```

Linear regression for percent of test takers with offers against borough

The Bronx is the intercept here and all other values are compared to the Bronx. Since all values are positive, a school in any other borough has a higher percent of test takers on average than a school in the Bronx.

A school being in Manhattan, for example, increases the percent of test takers with offers by about 24.67 compared to the Bronx.

Keep in mind that this regression output does not account for schools with fewer than five offers (see the 545 observations deleted noted in the output).

```
lm(`Percent of test takers with offers` ~ Borough, data = MS_data) %>% summary()
```

```
##
## Call:
## lm(formula = `Percent of test takers with offers` ~ Borough,
##     data = MS_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.562 -11.519  -0.458   6.935  46.206
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            14.787      5.583   2.649 0.009298 **
## BoroughBrooklyn         9.207      6.205   1.484 0.140793
## BoroughManhattan       24.674      6.384   3.865 0.000191 ***
## BoroughQueens          10.834      6.157   1.760 0.081303 .
## BoroughStatenIsland     2.427      8.172   0.297 0.767069
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.79 on 107 degrees of freedom
##   (545 observations deleted due to missingness)
## Multiple R-squared:  0.1905, Adjusted R-squared:  0.1603
## F-statistic: 6.296 on 4 and 107 DF,  p-value: 0.0001375
```