Report on demographics for “formated data.xlsx”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable (reference) | Undefined | R | I | S |
| Site (site = 0) | **0.5607** | **0.2119** | **0.05284** | **0.2307** |
| State (state = -1) | **0.3264** | **0.08582** | 0.01158 | **0.06011** |
| Gender (gender = 0) | **0.004** | **0.005** | 0.0003 | **0.006** |
| Age (no reference)\* | 0.0002 | **0.007** | 0.001 | **0.01** |

\*For Age, I used a quadratic model because of the pattern I noticed with the coefficients from the linear categorical model

This table shows the R2 values, **bold** represents a significant model (p < 0.001)

Have to adjust for multiple comparisons for site and state

All of the I results should basically be ignored since there are so many without I

Site:

I defined majorly significant as p > 0.001

For undefined, all the majorly noteworthy coefficients were negative, and were sites above 95)

(95, 98, 99, 101,104-111, 113, 115-123, 126-128, 130, 138, 140, 142, 145, 146,148, 149) All were on average about 2 lower than the rest.

For R, (16, 21,29,30,35,43,58,78,80,82,92,137) were the sites with majorly significant coefficients.

For S, (13,17,21,23,24,26,29,30,35,58,78,80,82,88,92,95,137,139) are the sites with majorly significant coefficients.

Age is related to S and R. From age categories 4-17 show an increase in R and a decrease in S compared to the groups on the ends, and a quadratic model is very significant, with a range of about 1.2 which represents a meaningful difference for a count.

The information about site is important to the Undefined / R / S values. Site is obviously tied to state so we do not need these correlated demographic variables, thus only site should be included.

Although the models were significant for gender, the coefficient for gender in these models were all less than 1 and not close to that value. (-0.14, 0.53, -0.56, for undefined, S, and R respectively) which suggests that even though a difference could exist it won’t really be meaningful in magnitude.

In conclusion, we should use site and age, while keeping in mind the quadratic relationship between age and the number of S and R’s.