Mini Case Study

Matthew O'Donnell

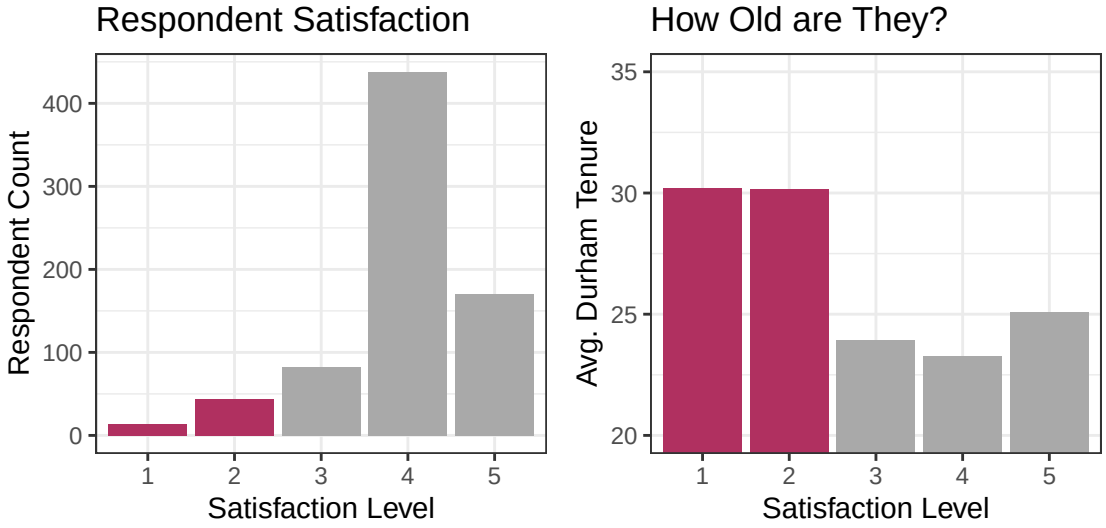
Introduction

The city of Durham has changed rapidly in recent years due to its increasing prominence as an educational hub. The city collects data annually to assess its residents' opinions on a variety of topics related to the state of the city. Due to the rapid changes in the city and influx of new residents, it is important to investigate the sentiment of Durham's longtime residents. This case study finds that long-term residents of Durham are less likely to be satisfied with the city as a place to live.

Methodology

The options for scoring Durham as a place to live are as follows:

1. Poor **2.** Below Average **3.** Neutral **4.** Good **5.** Excellent



The left-hand chart shows that the overwhelming majority of survey respondents report positive opinions of Durham as a place to live. To generalize this to the Durham population, it would require confidence that these respondents are representative of Durhamites.

The right-hand chart shows the average tenure in Durham for respondents who reported each satisfaction level. Respondents who reported lower levels of satisfaction had a longer average tenure, which is the first indication of a relationship. A standard linear regression or an ordinal regression would require that a similar trend in tenure continues across all satisfaction levels. However, in this chart, it is evident that the downward trend does not seem to continue between levels 3, 4, and 5. Therefore, a more suitable model might be a logistic regression where the outcome variable is a binary outcome of respondents who are satisfied with Durham (levels 3, 4, and 5) and unsatisfied (levels 1 and 2).

Various economic and demographic factors that may be significant predictors of satisfaction with Durham will also be included in the model (income, race, etc). This will reduce the likelihood of other confounders obscuring the true relationship between tenure and satisfaction.

This data is very incomplete, as many respondents only answer some of the questions. For Durham tenure and Durham satisfaction, missing responses will need to be removed. These are the main variables of interest, and missing values are not much use without complicated imputation. For any other variables that might be included in the model, missing categorical inputs can be retained as their own category, while missing numeric or ordinal inputs will need to be removed.

Model Equation

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 * tenure_i + \beta_2 * income_i + \beta_3 * housing_i + \beta_4 * own_i + \beta_5 * DPS_i + \sum_{j=6}^{9} \beta_j * race_{ij}$$

- p_i represents the probability that survey respondent i answered with a 3, 4, or 5 when asked to rate the Durham community as a place to live. Respondents who did not answer this question were removed.
- $tenure_i$ represents the length of time survey respondent i has lived in Durham. Respondents who did not answer this question were removed.
- $income_i$ represents which of four income brackets respondent i falls into. Respondents who did not answer this question were removed.
- $housing_i$ is the binary outcome of whether housing costs exceed 30% of monthly income for respondent i. Missing values take their own category.
- own_i represents whether respondent i owns their home. Missing values were assigned a 0 due to their small number.
- DPS_i represents whether respondent i currently has children in Durham public schools. Respondents who may have skipped this question would be assigned a 0.
- $race_{ij}$ represents whether respondent i indicated they identify as the j^{th} race. Respondents who skipped this question would have a 0 for all races. The effect for American Indian and Pacific Islander was not included due to the small number of respondents.

Results

Table 1: Model Estimates

term	estimate	p.value
(Intercept)	1.7688722	0.0153500
tenure	-0.0148478	0.0433694
income	0.3092972	0.0563746
housingNo Response	0.1863828	0.7765782
housingYes	-0.5801926	0.0655842
own	0.1182238	0.7138078
DPS	-0.0206499	0.9584385
Asian_Indian	0.2359063	0.7534998
Black	0.6880690	0.2359858
White	0.3623233	0.5043452
Latino_Hispanic	0.3531508	0.4430279

Discussion

With a $p < \alpha = 0.05$, the model results show a statistically significant negative relationship between tenure and satisfaction with Durham. In fact, it is the only model covariate with a statistical significant effect (although multicollinearity between inputs like *income*, *housing*, and *own* might be diluting each others' effects). The results tell us that for respondents who are identical across all other inputs (same income bracket, same race(s), etc.), each additional year of living in Durham decreases the estimated log-odds being satisfied with Durham by roughly 0.015.

The main implication of these results is that in a rapidly changing metropolitan area, those who have lived in Durham longer might be feeling more of the growing pains. In future urban planning, it will be important to include the voices of tenured residents in decision-making.

However, since fewer than 8% of all respondents reported that they were unsatisfied, it is worth questioning the overall importance of this finding. It is possible that by arbitrarily binarizing this highly imbalanced dataset to best match the relationship between tenure and satisfaction, it could increase the chances of indentifying a spurious relationship as significant. To better confirm the results, a larger sample of data could be collected with more complete responses.

An example of strange results that can come from this imbalanced dataset is that all races had a positive coefficient estimate. This suggests that belonging to any race increases the estimated probability of satisfaction, which seems counterintuitive since all people belong to at least one race. The root of this phenomena is that small group of unsatisfied respondents in this sample were less likely to report any race at all. This seems like a questionable result, and it emphasizes the need to confirm the model's results over a larger sample size.