Predicting the Number of Telecom Towers by Zip-Code Using Census Demographic Data

Alex Zharchuk

STAT 632

Introduction:
Cell networks are a revolutionary technology bound to disrupt the way we work, compute tasks, and handle information - especially as we upgrade our telecommunication systems to 5G and onwards.

Unfortunately, there has been no conclusive evidence of this technology's safety in close proximity to biological organisms.[1] The US also has a reputation for poor legislation regarding corporations and public health.[2] Thus, as a combination of indeterminate studies and mistrust of government transparency, it can be of concern that telecommunication systems emissions are a potential public health risk.

In order to minimize public health risks, there must be an examination of data pertaining to specific demographics and the response variable at hand - in this case the number of telecommunication towers. Accordingly, this paper introduces a statistical model and analysis pertaining to the relationship between the number of telecommunication towers and demographic data.

Data Description:
The data involved in this analysis is markedly quantitative. The predictor variable - number of towers - is quantitative, as is median household income, percent poverty, number of households, and percent disabled under 65. There is one binary variable and one categorical variable, respectively: presence of at least one airport (presence of airport = 1, no airport = 0), socioeconomic status. All so far listed variables but the response variable (number of towers) are predictor variables.

The data was sourced and compiled manually using three data sources: U.S. Census zip-code data, and two Google Earth layers from FCCinfo.com and the U.S. Census. For zip-code delineation within Google Earth, cartographic boundary files in KML format were obtained from census.gov.[3] Likewise, FCCinfo.com provided a KMZ file containing the locations of all operational transmitters registered to the FCC's Antenna Structure Registration database.[4]

Using Google Earth layers containing zip-code and transmitter locations, data-points for ASR tower locations were assessed and tallied visually within Google Earth, recorded into a 8✕30 .csv file via Google Sheets (Figure 1). Addedly, 30 zip-codes were selected off of various locations in California - mostly the Bay Area.

The sampling procedures for selecting zip-codes were not random, and were characterized by convenience sampling methods. Zip-codes were solicited from students within Discord and were also visually and preferentially chosen by the tallier.

A brief analysis of the data projected summary statistics was performed on each of the predictors and the response variable. The mean and standard deviation of variables of interest includes 6.03± 6.34 towers, household income of \$103,493.5 ± \$42,038.86, percent poverty of 11.56% ± 5.95%, number of households of 14,021.77 ± 4,783.06, and number of airports around 0.23 ± 0.01 in a given area among others (Table 1).

Methods:
In this study, a sample of 30 zip codes were used to investigate the relationship between various predictor variables and the number of towers in a given area. Before fitting our model, collinearity was tested by calculating the correlation between each predictor variable. The only variable that had significant correlation was poverty and median income of households (Figure 8), which makes sense because poverty is defined as income below federal poverty threshold levels.

We conducted a multiple linear regression analysis to examine the relationship between the number of airports and selected quantitative predictors in a given zip code. Collinearity was checked by testing correlation between each predictor variable. A full model that included all predictors of interest was examined for statistical significance. Then, a stepwise model selection was performed to select the most important predictors in our model, as well as to reduce overfitting. We then proceeded to check the assumptions of linearity, normality and homoscedasticity using both graphical and formal tests (ie. residuals vs fitted plot, QQ plot, Shapiro-Wilk's and Breusch-Pagan test). Using Box-cox transformation, we identified the type of transformations required to stabilize any assumptions violated. Multiple models were evaluated to determine the best one for our research question.

Results:
Fitting our model revealed that in predicting the number of towers, the statistically significant predictor was the number of airports in a given area (Figure 2). The stepwise model selection chose the total number of households and total number of airports as the two predictor variables in our reduced model. The AIC was 107.03 and we observed that for every additional household and airport in the area, the number of towers is estimated to increase by 0.0003299 and 6.8903791 towers respectively. The resulting model is as follows:

$\widehat{Number\ of\ Towers}$ = -0.2007480 + 0.0003299(Total Number of Households) + 6.8903791(Total Number of Airports)

This reduced model was also tested against the full model by the R function anova() to see if there was a statistically significant difference between the reduced model and the full model. The result was that there was not a statistically significant difference between the reduced and full model. In the interest of parsimony and the backing of our statistical tests we proceeded with the reduced model. We proceed to check the assumptions, and observe a violation in normality, linearity and homoscedasticity (Figure 3-4). A Shapiro-Wilk's test was performed to corroborate the graphical observations, and the resulting p-value of 0.0002 led us to conclude that the normality assumption is violated.

An attempt to improve the assumptions was done through a Box-Cox transformation where it was found that the estimated power of lambda was 0. This indicates that a log transformation was necessary. Multiple combinations of log transformations were performed in our model. A few outliers were also observed and its removal from the dataset was evaluated with the reduced and transformed model. Upon comparing and rechecking the assumptions with the models, assumptions were observed to have drastically improved in the transformed model as seen graphically (Figure 6). Performing a formal test with the Shapiro-Wilk's test resulted in a p-value of 0.1029, so we fail to reject the null hypothesis, and conclude that the normality assumption was met. Furthermore, given normality has been met, we can utilize the Breusch-Pagan test to evaluate homoscedasticity. Fortunately, we observe a p-value of 0.8757, and can conclude that the homoscedasticity was met. The resulting model is as follows:

$log(\widehat{Number\ of\ Towers})$ = -0.5 + 0.0001225(Total Number of Households) + 1.275(Total Number of Airports)

Furthermore, the regression summary table (Figure 7) indicates that both predictors are statistically significant after applying transformations and removing outliers, resulting in an improved model with an adjusted R-squared of 0.485.

Conclusion:

The study used multiple linear regression analysis to explore the relationship between the number of telecommunication towers and several quantitative predictors. The best model achieved an adjusted R-squared value of 0.485, which is relatively on the lower side. As a result, even with the improved R-squared value, we continue to be cautious about using the model as a predictive tool. It is worth noting that this value was obtained after removing outliers, which reduced the total number of data points to only 22.

Nevertheless, the selected predictors in this study included the number of households, and the number of airports in the given zip code. The analysis revealed that both the total number of households and the presence of an airport had a positive relationship with the number of telecommunication towers. Surprisingly, the analysis did not find that poverty or the percentage of population with a disability significantly predicted the number of telecommunication towers in a given area. However, it should be noted that this analysis only explored a limited set of predictors and that there may be other important factors that influence the distribution of telecommunication towers in a given area.

Moreover, a major limitation of this study is the small sample size, which may affect the generalizability of the findings to the actual population. Additionally, convenience sampling was used, which could introduce sampling bias and limit the representativeness of the sample. To address these limitations, future research could consider using a random sampling method to increase the sample size and reduce the potential for bias. We could also include zip codes in other states, and zip codes that aren't concentrated in and around the San Francisco Bay area in order to increase the generalizability of this model.

Furthermore, counting towers manually is a tedious and time-consuming task that may be prone to errors. Thus, a potential future direction for this research could be to develop a mechanism by which Google Earth layer information is parsable by automation. This could greatly improve the efficiency and accuracy of data collection and analysis.

Code Appendix: https://github.com/jamiecalma/telecom-towers

Figures:

Figure 1:

Visual data collection method in Google Earth. Zip-code delineation (left), vs. zip-code delineation and tower locations (right).

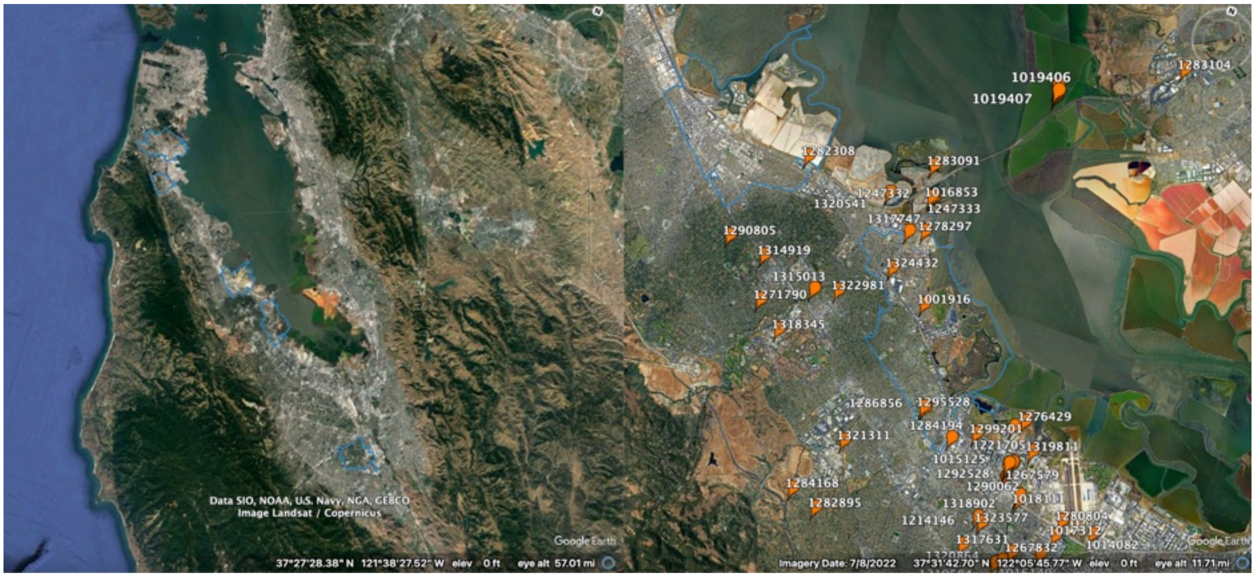
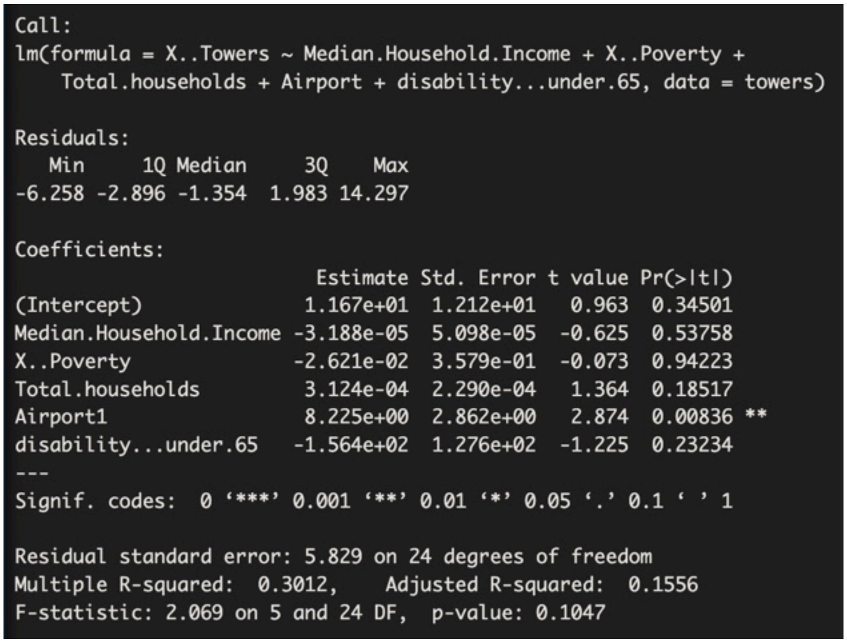


Figure 2: *Full linear regression model*



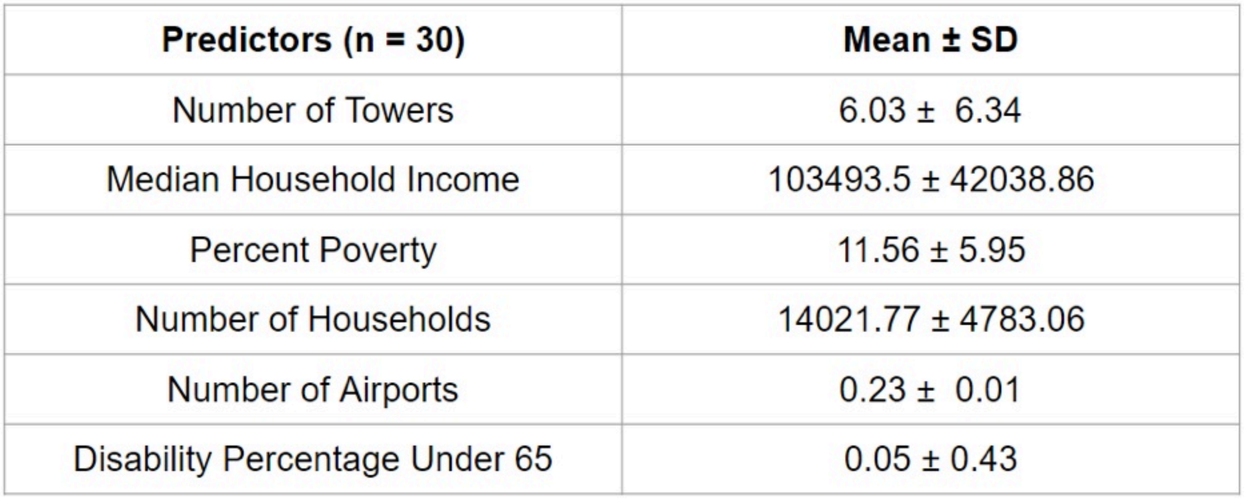
```
Call:
lm(formula = X..Towers ~ Median.Household.Income + X..Poverty +
    Total.households + Airport + disability...under.65, data = towers)

Residuals:
   Min     1Q Median     3Q    Max
-6.258 -2.896 -1.354  1.983 14.297

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.167e+01  1.212e+01   0.963  0.34501
Median.Household.Income -3.188e-05  5.098e-05  -0.625  0.53758
X..Poverty              -2.621e-02  3.579e-01  -0.073  0.94223
Total.households         3.124e-04  2.290e-04   1.364  0.18517
Airport1                 8.225e+00  2.862e+00   2.874  0.00836 **
disability...under.65   -1.564e+02  1.276e+02  -1.225  0.23234
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.829 on 24 degrees of freedom
Multiple R-squared:  0.3012,    Adjusted R-squared:  0.1556
F-statistic: 2.069 on 5 and 24 DF,  p-value: 0.1047
```

Table 1*: Summary statistics of data.*



Predictors (n = 30)	**Mean ± SD**
Number of Towers	6.03 ± 6.34
Median Household Income	103493.5 ± 42038.86
Percent Poverty	11.56 ± 5.95
Number of Households	14021.77 ± 4783.06
Number of Airports	0.23 ± 0.01
Disability Percentage Under 65	0.05 ± 0.43

Figure 3: *Scatterplot produced by pairs() from base R language*

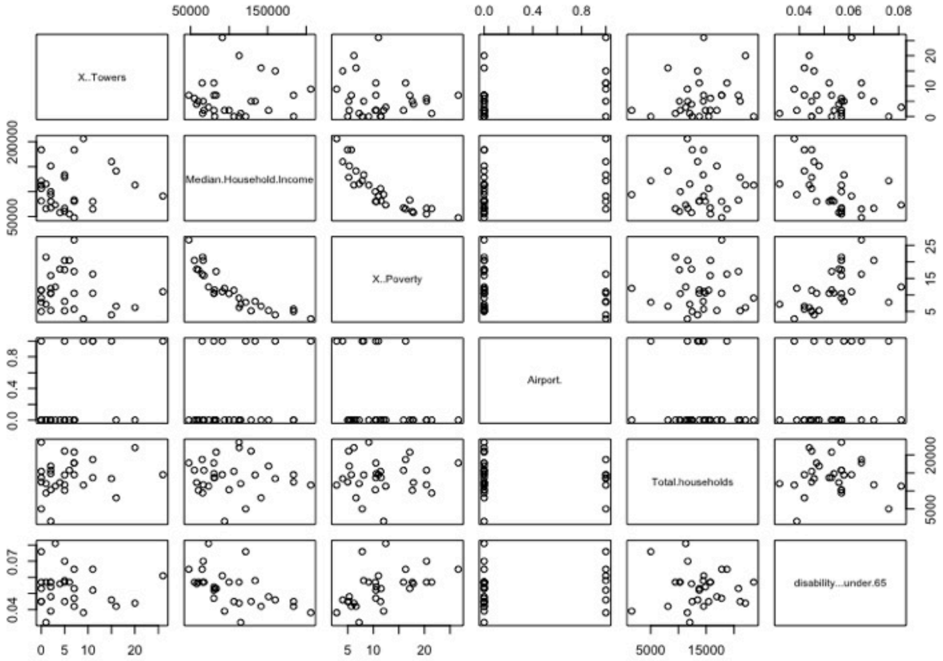


Figure 4: *Diagnostic plots for reduced model*

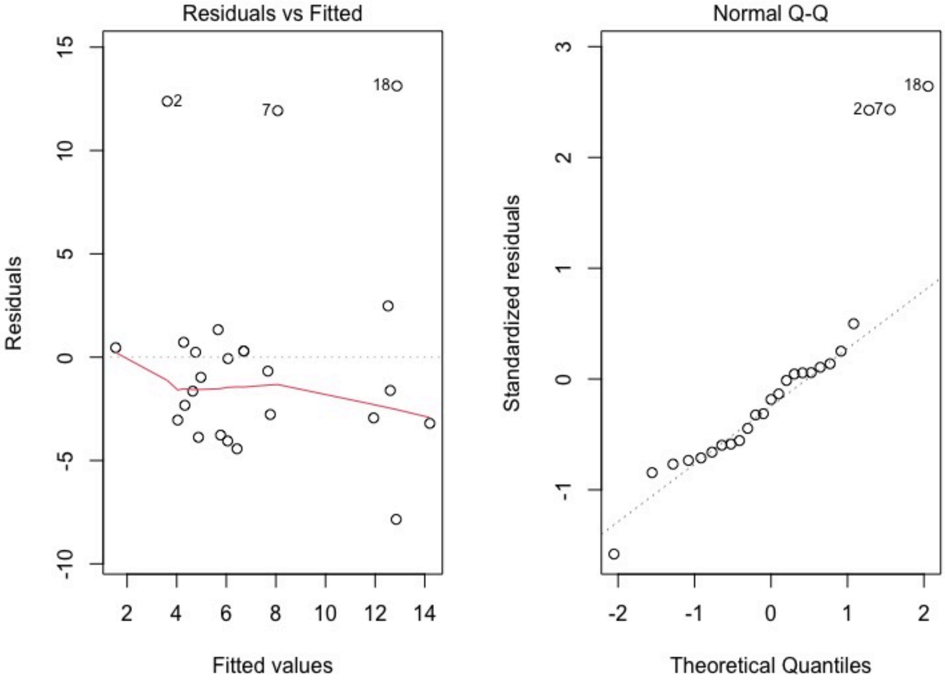


Figure 5: *Box-Cox Transformation Plot*

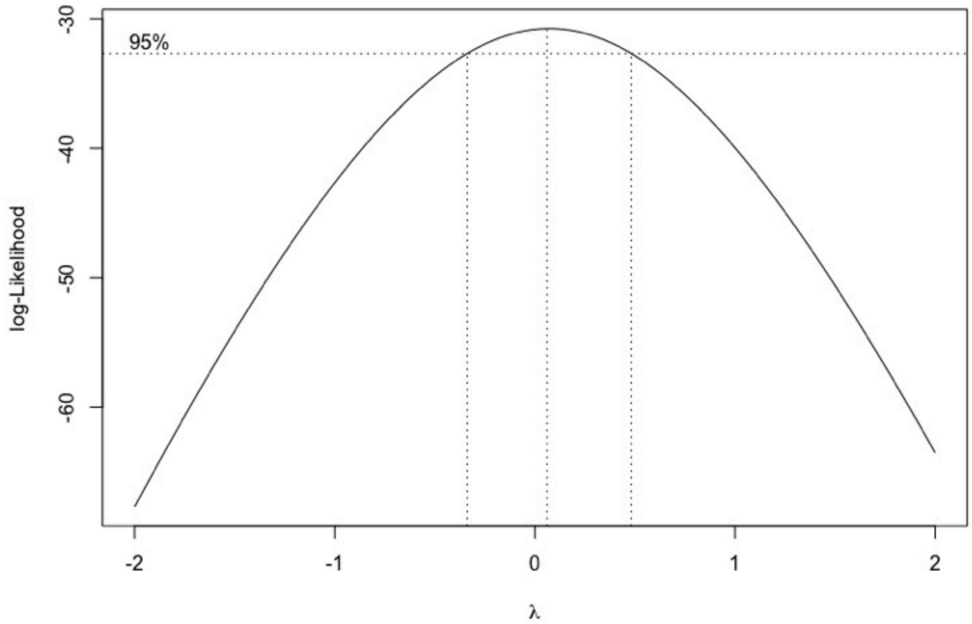


Figure 6: *Diagnostic plots for transformed model*

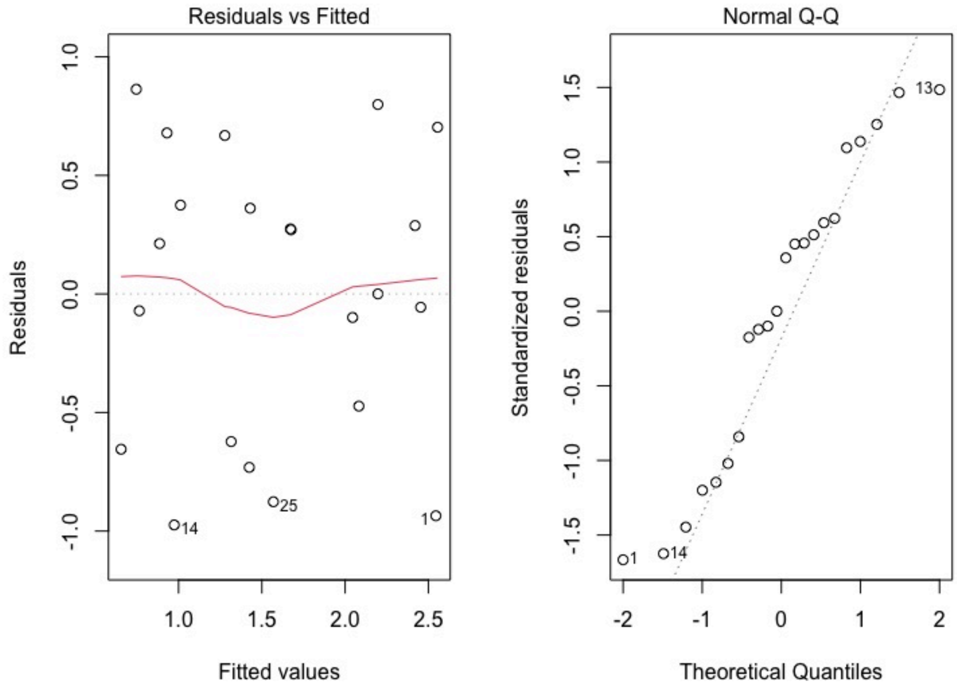
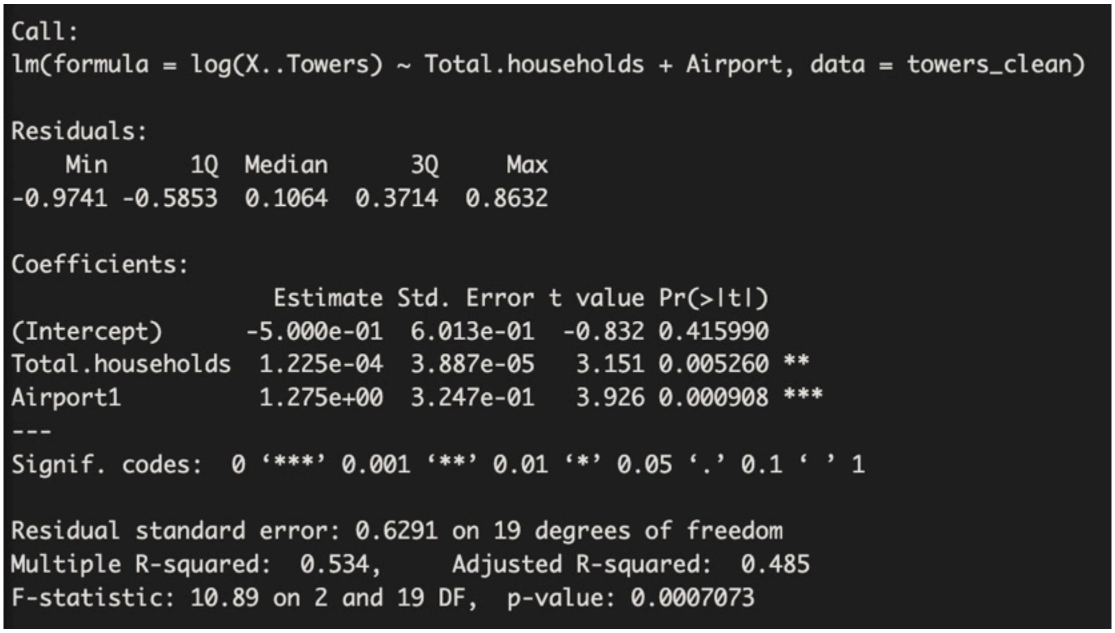


Figure 7: *Regression Summary Table for transformed model*



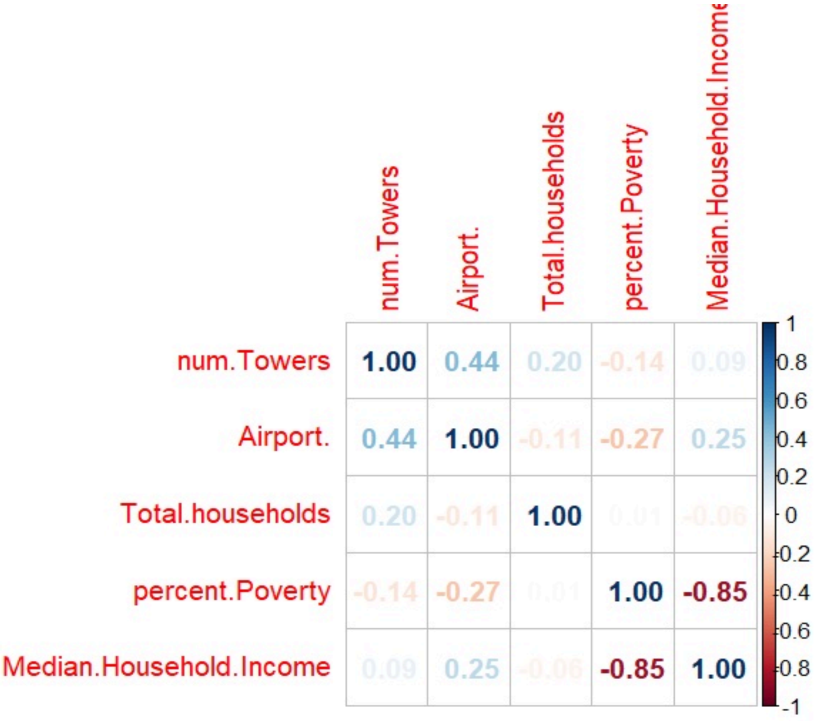
```
Call:
lm(formula = log(X..Towers) ~ Total.households + Airport, data = towers_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9741 -0.5853  0.1064  0.3714  0.8632 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -5.000e-01  6.013e-01  -0.832 0.415990    
Total.households  1.225e-04  3.887e-05   3.151 0.005260 ** 
Airport1          1.275e+00  3.247e-01   3.926 0.000908 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6291 on 19 degrees of freedom
Multiple R-squared:  0.534,	Adjusted R-squared:  0.485 
F-statistic: 10.89 on 2 and 19 DF,  p-value: 0.0007073
```

Figure 8: *Collinearity matrix of variables*



	num.Towers	Airport.	Total.households	percent.Poverty	Median.Household.Income
num.Towers	1.00	0.44	0.20	-0.14	0.09
Airport.	0.44	1.00	-0.11	-0.27	0.25
Total.households	0.20	-0.11	1.00	0.01	-0.06
percent.Poverty	-0.14	-0.27	0.01	1.00	-0.85
Median.Household.Income	0.09	0.25	-0.06	-0.85	1.00

References:

1. Moskowitz, Joel M. “Effects of Exposure to Electromagnetic Fields: Thirty Years of Research.” *Effects of Exposure to Electromagnetic Fields: Thirty Years of Research*, 4 Feb. 2019, www.saferemr.com/2018/02/effects-of-exposure-to-electromagnetic.html.

2. Shapiro, Joseph S. *Pollution Trends and US Environmental Policy: Lessons from the Last Half Century*. National Bureau of Economic Research, 2021.

3. US Census Bureau. “Cartographic Boundary Files - KML.” *Census.Gov*, 11 July 2022, www.census.gov/geographies/mapping-files/time-series/geo/kml-cartographic-boundary-files.html

4. Cavell, Mertz & Associates. “Now on Google Earth!” *FCCInfo*, www.fccinfo.com/fccinfo_google_earth.php . Accessed 8 April 2023.

5. Layer Coordinate Filtering Criteria, Antenna Structure Registration system https://wireless2.fcc.gov/UlsApp/AsrSearch/asrRegistrationSearch.jsp

6. Shapiro, Joseph S. “Pollution Trends and US Environmental Policy: Lessons from the Past ...” *The University of Chicago Press Journal*, www.journals.uchicago.edu/doi/10.1086/718054. Accessed 8 May 2023.