Business Problem

When male-male couple choose to relocate, among all the KPIs in this research, **cns_ratemm** - the density of male-male households is the top concern. We want to predict **cns_ratemm** using other practically available variables.

Linear Regression Model

cns_ratemm is our response variable here. The tax filing related variables can be calculated. (https://www.taxpolicycenter.org/sites/default/files/publication/153351/samesex_married_tax_filers_after_windsor_and_obergefell.pdf). The tax related variables, especially **taxrate_mm**, will be considered as the main explanatory variables.

I built a linear regression model using a cleaned dataset with 1302 rows of LGBTQ communities in 1302 different zip codes. I used **cns_ratemm** as the response variable and other useful information as explanatory variables. The model formula is

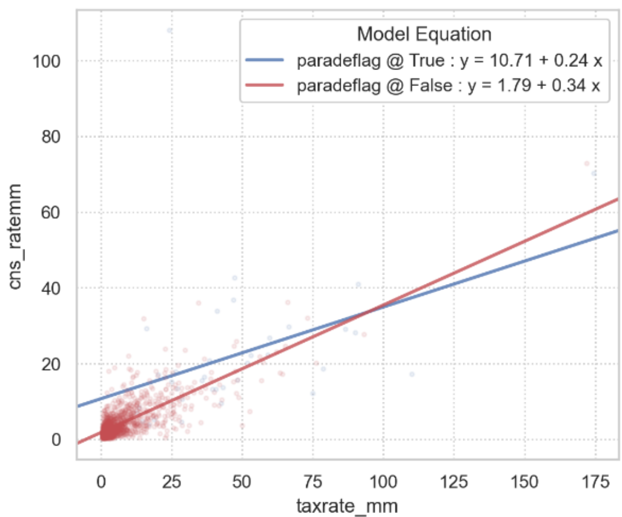
cns_ratemm ~ taxrate_mm + paradeflag + taxrate_mm:paradeflag

This model has 4 parameters with R^2 = 0.57. See Appendix 1 for model summary and estimated coefficients.

Model:

$$\begin{aligned}(\text{trip seconds}) &= 1.79 + 0.34(\text{taxrate mm}) + 8.92 \times \mathbf{1}_{\{\text{paradeflag}\}} - 0.09(\text{taxrate mm})\mathbf{1}_{\{\text{paradeflag}\}} \\ &= (1.79 + 8.92 \times \mathbf{1}_{\{\text{paradeflag}\}}) + (0.34 - 0.09 \times \mathbf{1}_{\{\text{paradeflag}\}})(\text{taxrate mm})\end{aligned}$$

R^2 = 0.57



Key Considerations in Modeling

- Look at all other columns' correlations to the column **cns_ratemm** and rank the correlations. **totindex**, **cns_upmm** and **taxrate_mm** rank the highest.

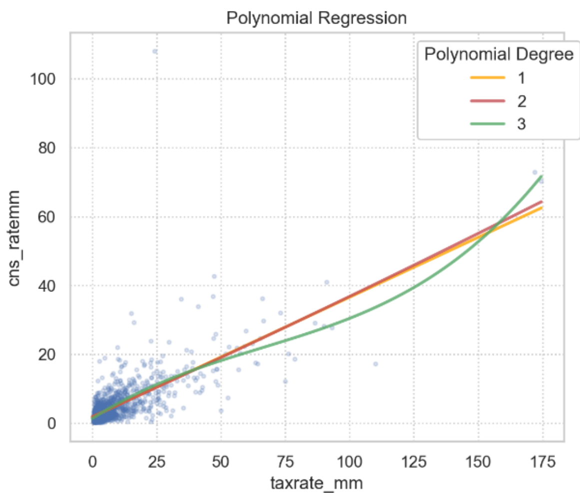
	tax_mjoint	geoid10	cns_tothh	cns_rateff	cns_upff	mjoint_ff	paradeflag	taxrate_ff	countbars	mjoint_mm	taxrate_mm	cns_upmm	totindex	cns_ratemm
cns_ratemm	-0.031871	0.06332	0.109584	0.202493	0.285508	0.316029	0.404067	0.407464	0.482626	0.590491	0.674153	0.688542	0.723054	1.0

- **totindex** is calculated from all other columns so it is impractical to use totindex as an explanatory variable even though it is highly correlated with response variable cns_ratemm.
- cns_upmm/cns_tothh = cns_ratemm. cns_ratemm is calculated from cns_upmm. Even though **cns_upmm** is highly correlated to cns_ratemm, practically, it is hard to get the data for cns_upmm in this real business case.
- **taxrate_mm**: the density of MM couples who file taxes among all filers. It turns out to be the most correlated column to cns_ratemm among all tax related columns.

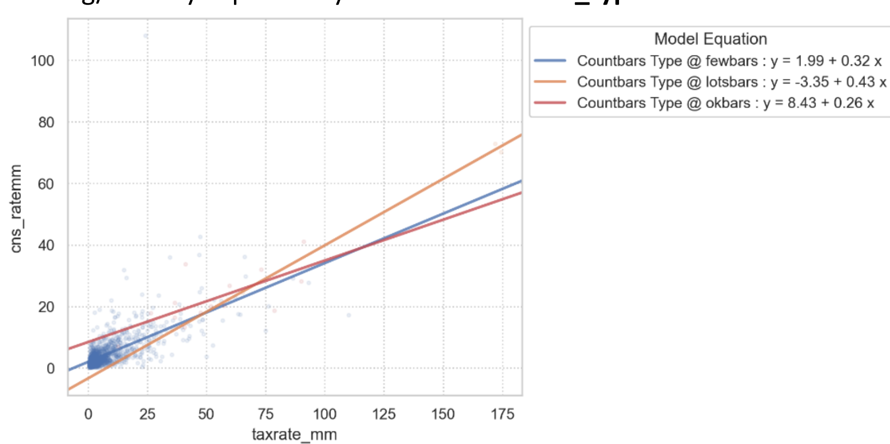
- If **geoid10** (zip code, unique identifier for each row in the dataset) is used as an explanatory variable, then it would surely exhaust the number of degrees of freedom of the model and yield 0 residual error i.e. overfit.
- **paradeflag** is transformed into a categorical column with 1 meaning pride parage goes through the zip code and 0 meaning pride parage does not go through the zip code.
- **countbars** is engineered into a categorical column **countbars_type** with valules of lotsbars, okbars and fewbars presenting if the zip code has lots, not many, or few gay bars.

Steps in Modeling

- **cns_ratemm ~ taxrate_mm**. Started with simple linear regression on the most practically impactful variable taxrate_mm. Resulted in R^2 = 0.55. This is treated as the baseline to compare when we add more variables later.
- **cns_ratemm ~ taxrate_mm + mjoint_mm.** Tried to include mjoint_mm, the next most correlated variable after taxrate_mm. Poor model explainability though higher R^2 because taxrate_mm and mjoint_mm are highly correlated. So taxrate_mm is used and mjoint_mm is excluded. Also, mjoint_mm and taxrate_mm both have high VIF. When all other explanatory variables exist in the model, mjoint_mm is redundant.
- **cns_ratemm ~ taxrate_mm + I(taxrate_mm ** 2) + I(taxrate_mm ** 3).** R^2 = 0.56. Tried cubic on **taxrate_mm** and got "The condition number is large, 3.27e+05. This might indicate that there are strong multicollinearity or other numerical problems." Polynomial was not adopted.



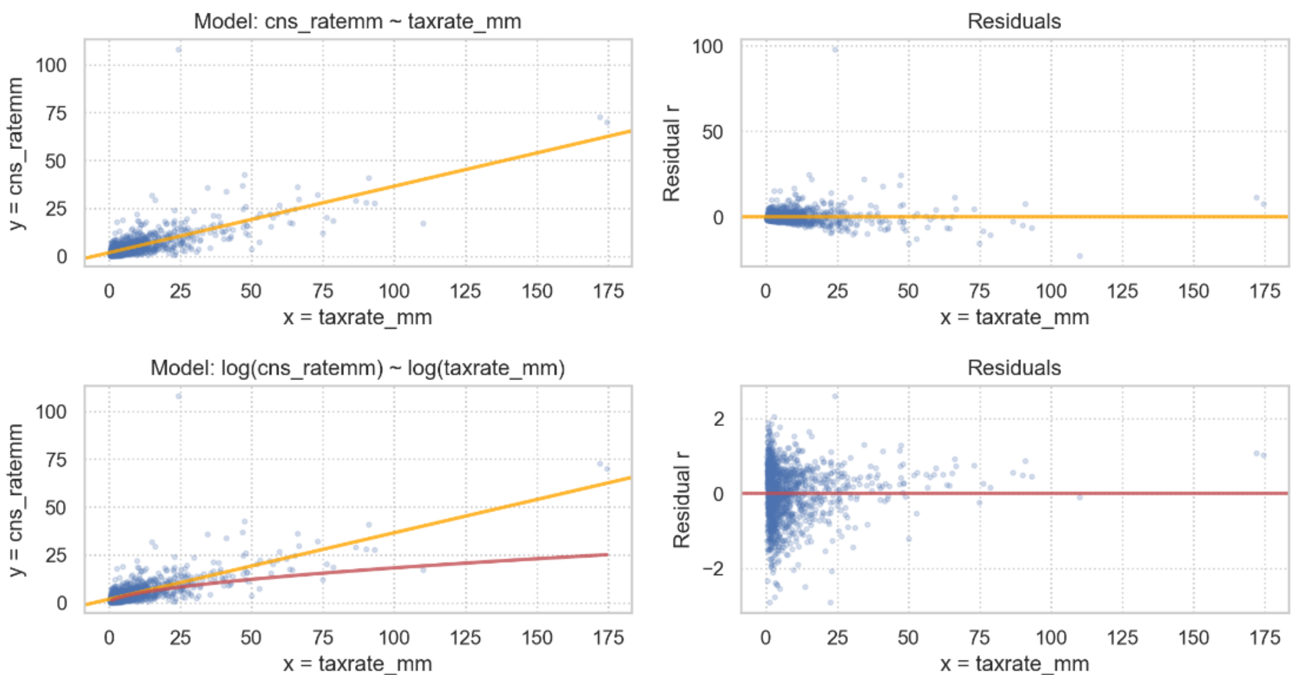
- **cns_ratemm ~ taxrate_mm + paradeflag**. Considering binary categorical explanatory variable **paradeflag.**
- **cns_ratemm ~ taxrate_mm + countbars_type**. Considering one-hot encoding/dummy explanatory variable **countbars_type**
- **cns_ratemm ~ taxrate_mm + paradeflag + taxrate_mm:paradeflag.** Considering binary categorical explanatory variable **paradeflag** and its interaction with **taxrate_mm.**
- **cns_ratemm ~ taxrate_mm + countbars_type + taxrate_mm: countbars_type**. Considering one-hot encoding/dummy explanatory variable **countbars_type** and its interaction with **taxrate_mm.**



- Table below summaries R^2 for the 4 regressions above. **cns_ratemm ~ taxrate_mm + paradeflag + taxrate_mm:paradeflag** is adopted with $R^2 = 0.57$.

cns_ratemm ~ taxrate_mm +		
R^2	no interaction w/ taxrate_mm	interaction w/ taxrate_mm
paradeflag	0.56	0.57
countbars_type	0.55	0.56

- There are some other numerical variables that we want to consider. The remaining outstanding ones that can be practically calculated are **taxrate_ff**, **cns_tothh** and **tax_mjoint**. Performed model selection with F-test. These extra variables passed the F test. But caused multicollinearity issues. "The condition number is large, 1.57e+05. This might indicate that there are strong multicollinearity or other numerical problems".
- Model diagnostic for heteroscedasticity. **np.log(cns_ratemm) ~ np.log(taxrate_mm)** with $R^2 = 0.46$ and decreases homoscedasticity.



- After considering steps above, **cns_ratemm ~ taxrate_mm + paradeflag + taxrate_mm:paradeflag** is adopted with $R^2 = 0.57$ for this business case.
- In this business case, both explanatory variables **taxrate_mm** and **paradeflag** are accessible and calculable. Response variable **cns_ratemm** can thus be predicted.
- This cleaned and feature engineered data set has 1302 rows, which can be a short-coming, especially for the categorical variables – **paradeflag** and **countbars_type**. Larger dataset would be ideal to tell paradeflag and countbars_type's effects on cns_ratemm more accurately.
- The process above adopted the forward regression method. The business goal is explicitly set to predict **cns_ratemm** based on other accessible and calculable variables. Backward regression is not suited here.

Appendix 1

```
print(smf.ols(formula='cns_ratemm ~ taxrate_mm + paradeflag + taxrate_mm:paradeflag', data=dfnozero).fit().summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             cns_ratemm   R-squared:                       0.569
Model:                            OLS   Adj. R-squared:                  0.568
Method:                 Least Squares   F-statistic:                     571.0
Date:                Mon, 18 Dec 2023   Prob (F-statistic):          1.52e-236
Time:                        00:33:38   Log-Likelihood:                -3726.0
No. Observations:                1302   AIC:                             7460.
Df Residuals:                    1298   BIC:                             7481.
Df Model:                           3
Covariance Type:            nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
Intercept                           1.7893      0.144     12.464      0.000       1.508       2.071
paradeflag[T.True]                  8.9245      1.229      7.260      0.000       6.513      11.336
taxrate_mm                          0.3365      0.011     30.803      0.000       0.315       0.358
taxrate_mm:paradeflag[T.True]      -0.0943      0.024     -3.932      0.000      -0.141      -0.047
==============================================================================
Omnibus:                     1939.862   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1564739.079
Skew:                           8.336   Prob(JB):                         0.00
Kurtosis:                     172.012   Cond. No.                         182.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```