0.1 Question 0: Human Context and Ethics

0.1.1 Question 0a

"How much is a house worth?" Who might be interested in an answer to this question? **Please list at least three different parties (people or organizations) and state whether each one has an interest in seeing the housing price to be high or low.**

Buyer: interested in seeing the housing price to be low. Seller: interested in seeing the housing price to be high. Tax Department: interested in seeing the housing price to be high.

0.1.2 Question 0b

Which of the following scenarios strike you as unfair and why? You can choose more than one. There is no single right answer, but you must explain your reasoning.

A. A homeowner whose home is assessed at a higher price than it would sell for.
B. A homeowner whose home is assessed at a lower price than it would sell for.
C. An assessment process that systematically overvalues inexpensive properties and undervalues expensive properties.
D. An assessment process that systematically undervalues inexpensive properties and overvalues expensive properties.

C would be an unfair situation because most of inexpensive properties attract lower income range of home buyer, and most of expensive properties attract higher income range of home buyer. If the first one is overvalued by the assessment process, meaning the home owner will have to pay more property taxes than those who buy undervalued expensive properties. In other words, this process will be biased towards the rich who get the undervalued expensive properties.

0.1.3 Question 0d

What were the central problems with the earlier property tax system in Cook County as reported by the Chicago Tribune ? And what were the primary causes of these problems? (Note: in addition to reading the paragraph above you will need to watch the lecture to answer this question)

One of the central problems with the earlier property tax system in Cook County was the houses in low-income neighborhood being overvalued while ones in wealthy community being undervalued, and redlinning as a result of both federal policy and rating system that takes races into account. The primary causes of these are having races as one the factors in the system, and similar patterns of a houses in a same community act as features in the model (e.g: neighborhood).

0.1.4 Question 0e

In addition to being regressive, how did the property tax system in Cook County place a disproportionate tax burden on non-white property owners?

In addition to the algorithms flaws, even homeowners were allowed to appeal their home values assesment, most of wealthy people who could afford lawyers tend to do this more than the poor. This results in the fact that the poor and non-white homeowners paid the property taxes more than the rich white homeowners.

0.2 Question 2a

Without running any calculation or code, complete the following statement by filling in the blank with one of the comparators below:

$$\geq$$
$$\leq$$
$$=$$

Suppose we quantify the loss on our linear models using MSE (Mean Squared Error). Consider the training loss of the 1st model and the training loss of the 2nd model. We are guaranteed that:

$$\text{Training Loss of the 1st Model} \geq \text{Training Loss of the 2nd Model}$$

In general, adding another featuren will cause the training error of our model decrease or stay the same. Thus, the first model with only one feature will alwats have training loss greater than or equal to training loss of the second model.

0.3 Question 3b

You should oberseve that θ_1 change from positive to negative when we introduce an additional feature in our 2nd model. Provide a reasoning why this may occur. **Hint:** which feature is more useful is predicting `Log Sale Price`?

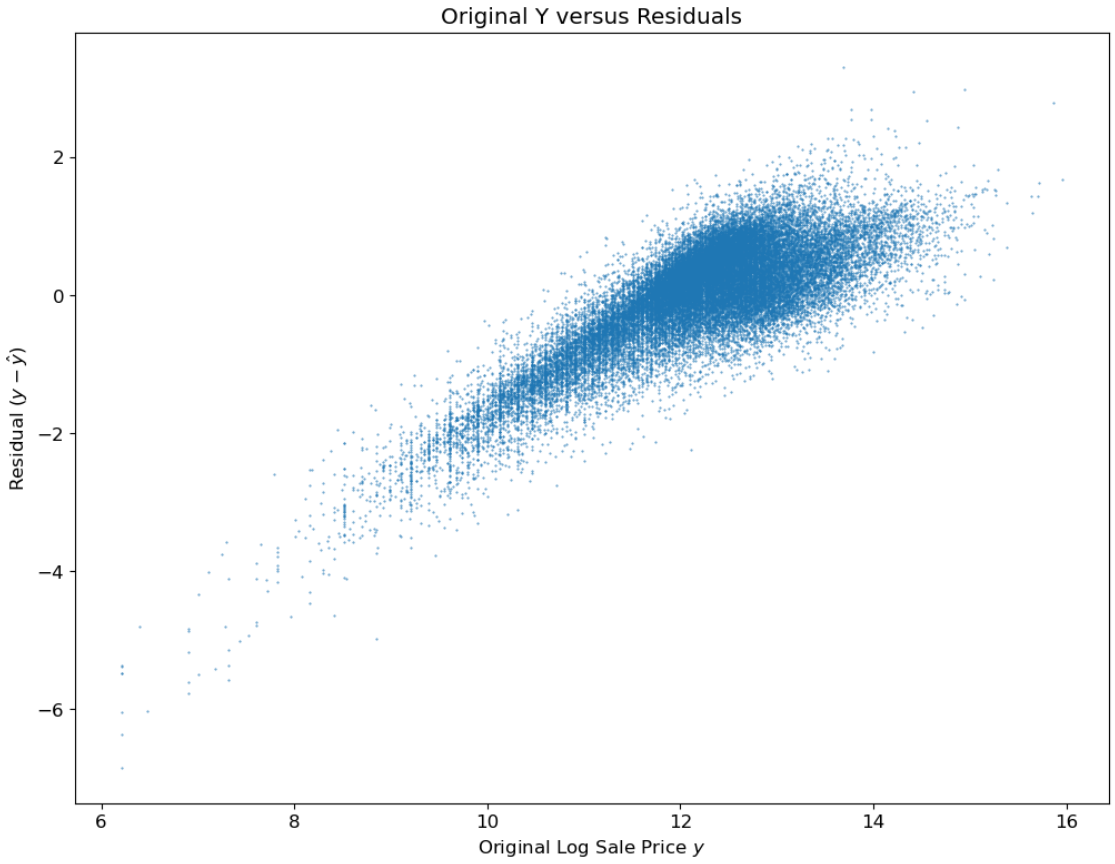
This may occure because the 2nd feature "Log BuildingSquare Feet" appears to be more helpful when predicting housing prices, thus our optimized model puts more weight on this feature over the first feature "Bedrooms". This explains why theta_1 becomes negative in model 2.

0.4 Question 3c

Another way of understanding the performance (and appropriateness) of a model is through a plot of the residuals versus the observations.

In the cell below, use `plt.scatter` to plot the residuals from predicting `Log Sale Price` using **only the 2nd model** against the original `Log Sale Price` for the **validation data**. With a data size this large, it is diffult to avoid overplotting entirely. You should also ensure that the dot size and opacity in the scatter plot are set appropriately to reduce the impact of overplotting as much as possible.

```
In [228]: plt.scatter(y_valid_m2, y_valid_m2 - y_predicted_m2, s=0.2,alpha=0.8)
          plt.xlabel("Original Log Sale Price $y$")
          plt.ylabel("Residual $(y - \hat{y})$")
          plt.title("Original Y versus Residuals");
```



0.5 Question 5

In building your model in question 4, what different models have you tried? What worked and what did not? Brief discuss your modeling process.

Note: We are not looking for a single correct answer. Explain what you did in question 4 and you will get point.

First, I chose all the numerical variables that their log transformations seem to have a linear association with log sale price, and hot-encoded all the categorical variables that their side-to-side boxplots also seem to have a linear association. Then, I cross-validated on 9 combinations of lower(500,510,520) and upper bound(710, 700, 690) for the column 'Sale Price' in model, and found out the optimal combination are at lower = 500 and upper = 710. After that, I checked my training and valid errors and saw that my model was overfitting because my training error is much lower than my valid error. So, I used Lasso model for my features selection. To see what is the optimal alpha for my Lasso model, I ran another cross validation on different alpha, and found out the optimal one is 0.002031. Then I checked the coefficient of Lasso model on that optimal hyperparameter alpha to see which features (or columns) I need to remove (coefficient = 0). That is my final model.

0.6 Question 6 Evaluating Model in Context

0.7 Question 6a

When evaluating your model, we used root mean squared error. In the context of estimating the value of houses, what does residual mean for an individual homeowner? How does it affect them in terms of property taxes? Discuss the cases where residual is positive and negative separately.

The formula for residuals are actual y - predicted y. To homeowners, if the residual is negative, their home is being overvalued, and they have to pay property taxes more than they are supposed to. On the other hand, if the residual is positive, their home is being undervalued, and they benefit from this by paying less property taxes than they should.

0.8 Question 6b

In your own words, describe how you would define fairness in property assessments and taxes.

Faireness in property assesments and taxes have to address whether the result will harm marginalized community by preventing regressive taxes over the model's accuracy. It has to take into account the wealth gap between neighborhoods to avoid using any features that could trace back to biased result such as races/ethnicity/ desiribility rate (that is determined by biased historical data).

0.9 Question 6c

Take a look at the Residential Automated Valuation Model files under the Models subgroup in the CCAO's GitLab. Without directly looking at any code, do you feel that the documentation sufficiently explains how the residential valuation model works? Which part(s) of the documentation might be difficult for nontechnical audiences to understand?

To some extent, I feel that the documentation gives a general idea of how the algorithm picks the features, and how the valuation process look like. However, this documentation is not sufficently explains to the public because it uses a lot of jargons that is difficult for non-technical audiences to fully understand. For example, sections like "Model Selection" and "Framework Selection" introduces "Tidymodels" or "LightGBM" without explaining how they works. If a person who doesn't come from machine learning community, they won't understand these complex vocabularies.