Exploring the Performance Impact of Approximately Decomposed Training Data

Chaoting Wu, Kuanyu Lu
Information School, University of Washington

Information School UNIVERSITY of WASHINGTON

Abstract

As previous works have shown, relational decomposition of training data enjoys both the efficiency of space and time while preserving the same performance under certain models (e.g. linear regression) and data conditions (Schleich et al., 2016). However, real-world data are often subject to noise, where exact decomposition as used in the previous work couldn't be obtained. Though this problem can be resolved by adopting approximate normalization (Kenig et al., 2019), it's unclear whether the performance of the model will be still preserved, assuming the speed-up holds true for approximate normalized schemes. In this project, the performance of the linear regression models obtained over approximate acyclic schemas, as a substitute for exact decompositions, is examined to explore the benefits or losses in addition to the assumed efficiency gains in real-world data-sets.

Methods and Materials

In this work, Maimon, an acyclic schema generator developed in Kenig et al., 2019, is used to explore possible approximate decompositions and schemas. Real-world datasets collected from Kaggle are used as our data source. For each dataset, the following steps are executed and repeated for many times under different random seeds:

- Original data will be split into a training set and a test set in a 80-20 split, where the training set is used for:
  - Relational decomposition by using Maimon under a given epsilon, which is the degree of approximation
  - Setting up the baseline model using the most naive full linear model
- The test set is used for assessing performance of the models obtained

We'll then examine if there's any correlation or phenomenon on the metrics that we might interested in, such as the percentage of spurious tuples against the performance of the model. Visualizations will be provided, analysis will be conducted, and we'll try to provide some possible intuitive explanations.

Experiments

We've conducted 5 experiments over the Red Wine Quality dataset[1] with 5 random seeds used for train-test split. For each experiment, we successfully generated around 30 - 32 acyclic schemas under epsilon = 0.1. We stored the decomposed tables in a PostgreSQL database and compute the joined result using views. For each joined result, we select the most representative linear model using a 5-fold cross-validation.

```
Acyclic scheme:
JMeasure: 0.1278637881605249
Num of JDs: 4
# clusters: 7
Max cluster size: 5
Max sep size: 3
 Data-intensive measurements:
# Spurious tuples: 141 total %: 0.8157361874457623
 Largest relation: 17280
 total tuples in decomposition : 120009
 total cells in decomposition : 497316
Printin Scheme # 309
separator, level: 1: {3, 4, 10}
cluster, level: 2: {1, 3, 4, 10}
cluster, level: 2: {3, 4, 6, 10}
separator, level: 2: {0, 3, 5, 10}
cluster, level: 3: {0, 3, 5, 7, 10}
separator, level: 3: {0, 4, 8, 10}
cluster, level: 4: {0, 4, 8, 9, 10}
cluster, level: 4: {0, 2, 3, 4, 5, 8, 10}
```

Results

164 acyclic schemas were generated in total and some schemas reported more than 45 times the number of the original rows. The models from original data reported an average MSE of 0.4 and R-Squared of 0.34, while the generated schemas reported an average MSE of 0.45 and R-Squared of 0.32.

Figure 1. Percentage of Spurious Tuples v.s MSE and R-Squared

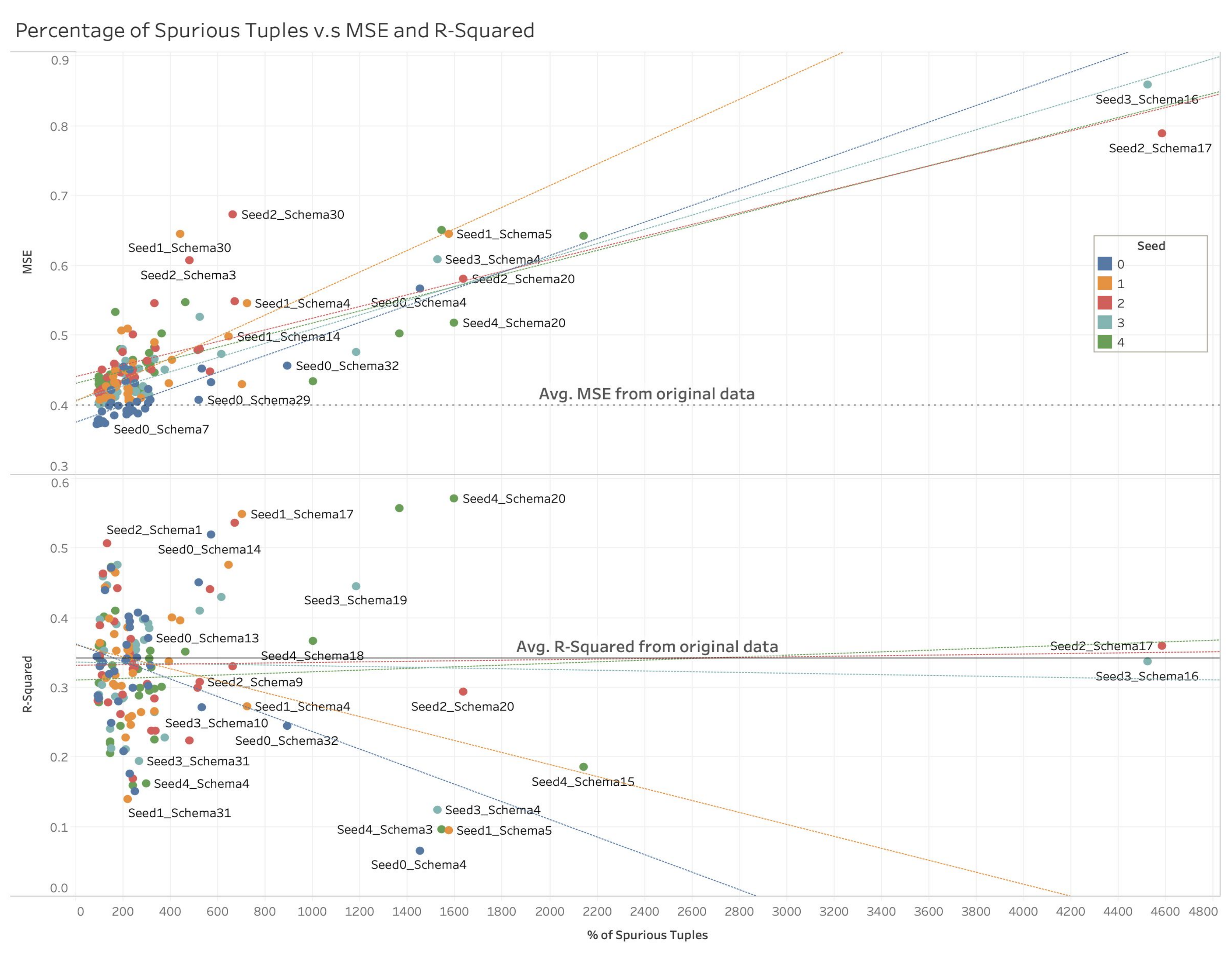


Figure 2. Residual Plots for selected schemas

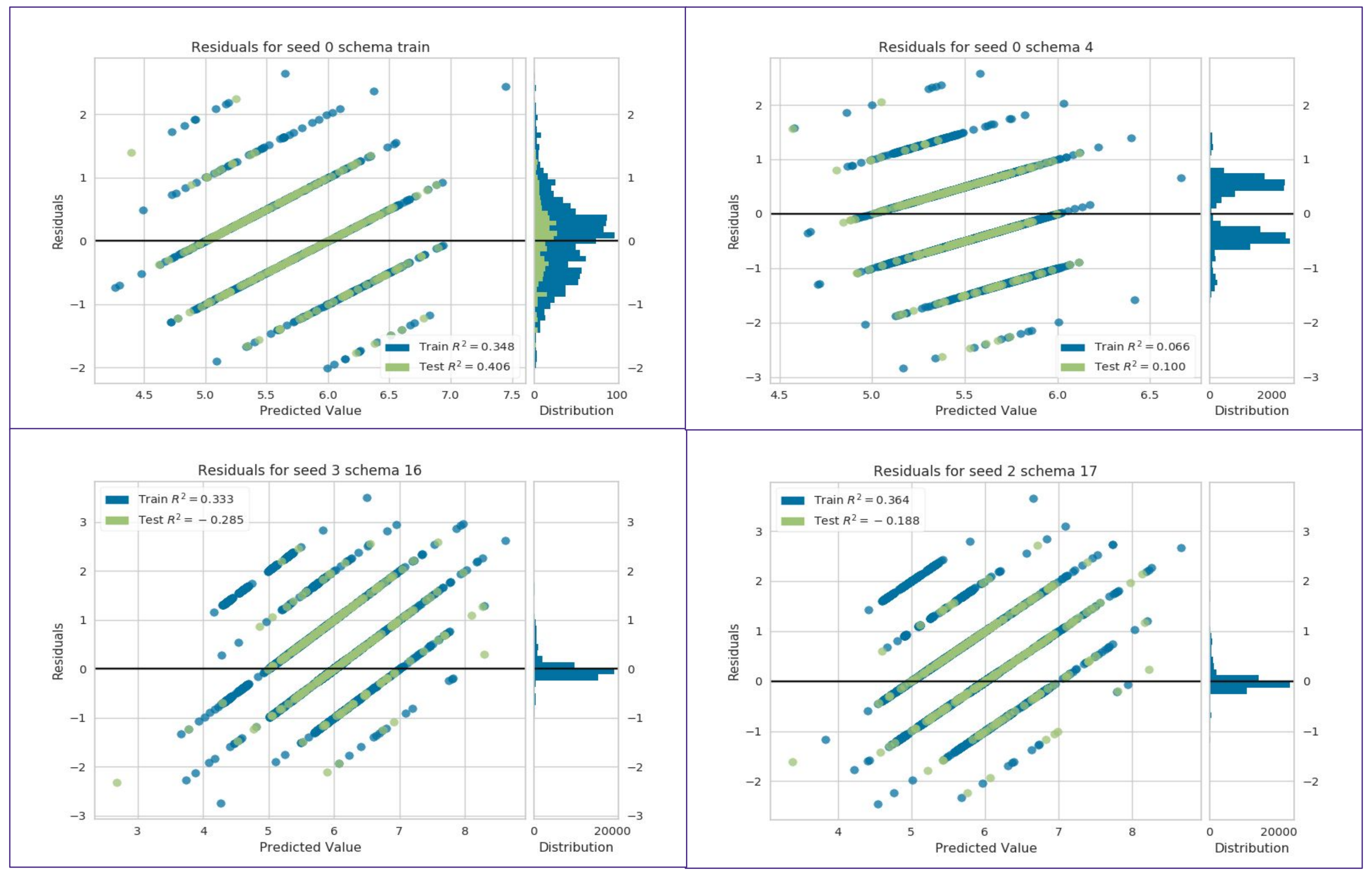
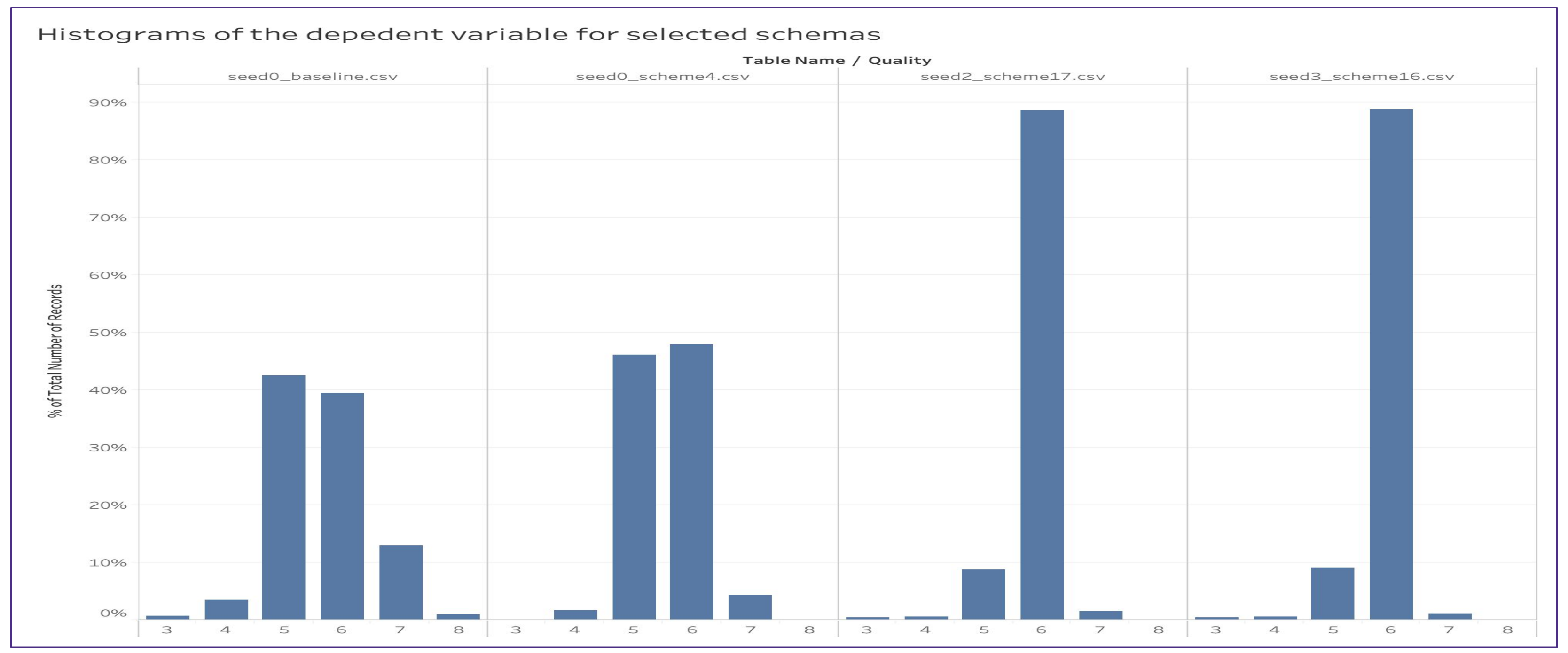


Figure 3. Histograms of the dependent variable for selected schemas



Analysis

Looking at Figure 1, there are two key takeaways here: (1) In general, more spurious tuples lead to lower performance (i.e. higher MSE) , and (2) more spurious tuples sometimes lead to lower R-squared value in the obtained models. To explain these 2 takeaways, residual analysis (Figure 2) is introduced. For takeaway one, by looking at the residual plots of schema 16 and 4 in experiment 2 and 3 respectively, which have the highest MSE over all the observations, we found out that's due to the change of the distribution of the dependent variable. In this two generated schemas, huge amounts of spurious tuples with popular dependent variable in the original data lead to distributions with high kurtosis (Figure 3). The models obtained from these two schemas thus have less explainability than the original model when predicted values are rare. For takeaway two, by looking at schema 4 in experiment (seed) 0, where we observed a slightly higher MSE but a extremely low R-Squared value (0.06), we found out that it's due to the fact that the data generated doesn't follow the assumption linear regression is based on, that residual should be normally distributed.

Conclusions and Future work

To conclude, we identified a few problems that decomposition on training data might have, if we choose to do so in order to leverage the efficiency it provides: (1) We might stumble upon some schemas that are distorted on the dependent variable, which will lead us to some really biased models, and (2) we could by chance get schemas that generate data completely different than the underlying assumptions we had on the original one, which in this case, the linear regression assumptions.

The King County House Price dataset, which has more than 12 columns and up to 20k rows is under experiments now. We're optimizing Maimon on this task since it seems too demanding for our machines. Hopefully we will derive similar results on this dataset with more convincing evidence, such as the residual plots since the dependent variable in this case is ratio data.

Contact

Chaoting Wu
iSchool, UW
chaoting@uw.edu

Kuanyu Lu
iSchool, UW
lukuanyu@uw.edu

References

1. Red Wine Quality, Simple and clean practice dataset for regression or classification modelling. https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
2. House Sales in King County, USA. https://www.kaggle.com/harlfoxem/housesalesprediction
3. Kenig, B., Mundra, P., Prasad, G., Salimi, B., Suciu, D..(2019). Mining Approximate Acyclic Schemes from Relations. arXiv.org. https://arxiv.org/abs/1911.12933
4. Schleich, M., Olteanu, D., Ciucanu, R..(2016). Learning Linear Regression Models over Factorized Joins. University of Oxford. http://www.cs.ox.ac.uk/dan.olteanu/papers/soc-sigmod16.pdf