Rental Price Modeling

An Exploration of Linear Models to Aid in Rental Price Predictions

Andrew Whittenbarger
Departments of Applied Mathematics, Computer Science, and Information Science
University of Colorado Boulder
Boulder, Colorado, US
andrew.whittenbarger@colorado.edu

ABSTRACT

This project explores two linear models and a traditional market averaging method with the intent of using data available at the end of the fiscal year, September, to project rental prices for the following calendar year. These models can serve as a tool to aid employers in creating a compensation budget at the beginning of the fiscal year that allows them to remain competitive in the following calendar year when adjusting compensation of current employees as well as recruiting new hires. Linear models appear to outperform traditional averaging by reducing error from 6.2% to 1.5%. However, 13% of linear models were found to be statistically insignificant at an alpha level of 0.05, which suggests future work is required to further refine linear models before widespread application is possible.

CCS CONCEPTS

•General and Reference •Cross Computing Tools and Techniques • Estimation

KEYWORDS

Fiscal Year (FY), Basic Allowance For Housing (BAH), Zillow Observed Rent Index (ZORI)

ACM Reference format:

Andrew Whittenbarger. 2022. Rental Price Modeling: An Exploration of Predictors to Aid in Rental Price Projections.

1 Introduction

Many companies set budgets at the end of the fiscal year in September. These budgets allocate a certain amount to employee compensation which is then used to retain and recruit employees well into the following calendar year. Housing costs contribute to over 20% of monthly expenses for most US citizens, and 1 in 5 households spend over 30% of their monthly budget on housing [1]. If an employer wishes to remain competitive, they must be aware of the cost of housing to employees and ensure they are offering compensation that allows for employees to secure adequate housing. The problem is that the cost of housing tends to be dynamic and often varies by market. Existing solutions are limited. If an employer sets a compensation budget in September based on an average of data collected from the current year, they are likely not capturing an accurate representation of what the cost of housing will be in the following year, especially during peak months where the housing market is the most competitive. Exploring different linear models can potentially help to better forecast the cost of housing the following calendar year.

2 Related Work

While many employers do not make their compensation models publicly available, the Department of Defense (DOD) has published their methodology on how they project housing costs and set the Basic Housing Allowance (BAH) for Military Members. Presently, the DOD collects data during the peak spring and summer months when the housing market is most active and uses these averages to determine the current market rent which is then used to set BAH rates for the following calendar year. However, it does not appear any predictors are used to forecast potential rent increases or decreases for the following year [2].

3 Proposed Work

The main task of this study is to build upon previous work by determining if using historical market rent averages with date as a linear model predictor helps forecast future rent values better than the current market average. It is important to note the scope of this project only explores rental prices. Home purchase prices will not be explored. While the Data Mining Series of classes focused mostly on discrete data, this will be a form of trend analysis that was touched on during the mining complex data lecture.

3.1 Dataset

The Zillow Observed Rent Index (ZORI) was used as the primary dataset for this project. The dataset is in aggregate on rental information collected on Zillow and affiliate companies websites as well as publicly available Census data [3]. When the dataset was downloaded on November 15th, 2022, the raw dataset consisted of average rent prices from March 2017 to October 2022. There were 47,151 data points, 507 metro areas, and 18,490 missing values. How missing values were addressed is discussed in detail in the main tasks section of the report. The summary of rent trends after data cleaning is given below.

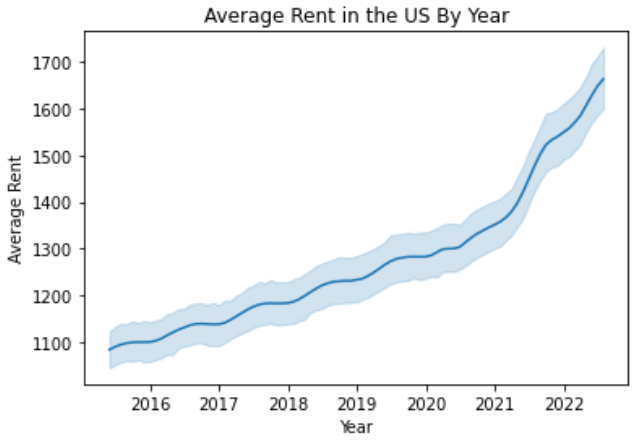

Figure 1: Average Rent Trends in the US

The average rent is represented by the dark blue line, while the 95% confidence interval is represented by the light blue shaded area. From mid 2015 to early 2021, it appears that rent increased in a linear fashion with a slightly cyclical pattern of rent increasing in the summer and decreasing in the winter. It also appears that rental increases in 2021 and 2022 were more significant than the previous years in the dataset.

3.2 **Tools**

Python was the programming language used to study the dataset. The Pandas library [4] was used during all steps of the data mining process, and the SciPy library [5] was used for linear modeling. The Seaborn library was used to visualize the data and create graphics [6].

3.3 **Main Tasks**

There were 3 main tasks associated with this study:

3.3.1 Data Preprocessing/Cleaning: The dataset was imported as a dataframe into Python. The data was cleaned by removing unused columns. For reasons discussed in the experimental setup dates prior to May 2015 and dates after July 2022 were trimmed from the dataset. Metro areas that were missing more than 5% of the remaining data points or had 2 consecutive months of missing data points were removed from the study. The remaining missing values were estimated as the average of the previous and post month's rent.

3.3.2 Data Warehousing: The original dataset assigned the dates as column headers which was not conducive to completing necessary calculations on the dataset. The data was transposed so that all dates are in a single column. A preview of the dataset used for modeling is given below.

	Metro	Date	Rent
0	Chicago, IL	2015-05-31	1447.457061
1	Chicago, IL	2015-06-30	1471.120201
2	Chicago, IL	2015-07-31	1477.492074
3	Chicago, IL	2015-08-31	1483.756181
4	Chicago, IL	2015-09-30	1481.774141

Figure 2: Refined Dataset Preview

3.3.3 Data Modeling: Linear models were built for each metro area with date as the predictor variable. The general linear model equation is given as [5]:

$$y = \beta_0 + \beta_1 * x \quad \textbf{(1)}$$

where:
$y = market\ rent$
$\beta_0 = rent\ at\ the\ model\ start\ date$
$\beta_1 = change\ in\ slope\ based\ one\ unit\ change\ in\ date$
$x = date$

The traditional average equation is given as:

$$y = \frac{y_{may} + y_{June} + y_{July}}{3} \quad \textbf{(2)}$$

where:
$y = market\ rent$
$y_{may} = may\ rent\ average$
$y_{June} = June\ rent\ average$
$y_{July} = July\ rent\ average$

3.3.4 Pattern Evaluation: Both models were compared to actual average rent for the peak summer months of the following year. The percent difference of models was calculated as:

$$pd = 100 * \frac{(PR-AR)}{AR} \quad (3)$$

where:
$pd = percent\ difference$
$PR = predicted\ rent\ (\$)$
$AR = average\ rent\ (\$)$

To avoid data dredging, a stepwise process was taken which is discussed at length in the experimental setup section. Evaluations are discussed in detail in the evaluation metrics section.

4 Experimental Setup

After data preprocessing and cleaning, there are 215 metro areas and 18,705 data points remaining in the dataset. 215 missing values were replaced with averages of the previous and post month rent. The experiment grouped the rental data by metro area and calculated the average “peak month” rent by year. For the purpose of this experiment, July data points are used as the peak month rent. Predictive values of the linear model and traditional averaging method are compared to the actual rent to assess effectiveness. To avoid data dredging, the calculations will be conducted in an iterative process as seen in table 1.

Table 1: Experimental Data Setup

Iteration	Max Input Date	Prediction Date
1	AUG 2015	JUL 2016
2	AUG 2016	JUL 2017
3	AUG 2017	JUL 2018
4	AUG 2018	JUL 2019
5	AUG 2019	JUL 2020
6	AUG 2020	JUL 2021
7	AUG 2021	JUL 2022

In the experimental setup the “Max Input Date” is the maximum date used to build the predictive model. The “Prediction Date” is the date for rent price predictions. 2 linear models were used in the comparison: a rolling predictive model which encompasses all available data up to the max input date and an annual predictive model that only uses data from the previous year. A total of 4,515 models were built during the experiment, 645 for each prediction year. In graphical representations the rolling model will be referred to as “lm1” and the annualized model will be referred to as “lm2”. The traditional averaging method will be referred to as the “Traditional”.

4.1 Evaluation Metrics

Results of each model were evaluated against the reported market average to determine effectiveness. Specifically, for each iteration the difference between the models and true mean were calculated as a percent difference as given in equation 3. Here, values closer to zero represent better performance since a zero value would represent no difference from the measured mean and prediction.

In order to determine if the difference between the traditional estimation method and the linear models were actually statistically different, a 95% confidence interval was calculated for the overall linear models and traditional averaging method as seen in table 2.

Table 2: 95% CI of Average Percent Difference

Model	Mean (%)	CI Lower (%)	CI Upper(%)
Traditional	-6.18	-6.41	-5.94
Rolling LM	-3.50	-3.80	-3.19
Annual LM	-1.13	-1.38	-0.89

The average percent difference for the models were in fact statistically different. The average percent difference for all models were below zero with the traditional averaging method being 6.18% off from the actual peak rent. In regards to percent difference, the best performing models ,on average, appeared to be the annualized linear model which was only 1.13% lower than the market rent. To put that in perspective, in August of 2022 the mean rent in the US was $2,495 per month [7]. A compensation plan built using the traditional average would undercompensate for rent by $1,850 per employee on an annualized basis, whereas a rolling linear model and annualized linear model would underpredict by $1,048 and $338 respectively.

A violin plot was also created to gain a better understanding of the distribution of each model's predictions. The median of each model is represented by a white dot while the interquartile range is represented by the

black box. The thin black lines are drawn to 1.5 times the interquartile range. Colored areas represent a kernel density estimation. Given 2 points on the y-axis, the colored area under the curve represents the comparative probability that a new random sample would fall between those 2 values [6].

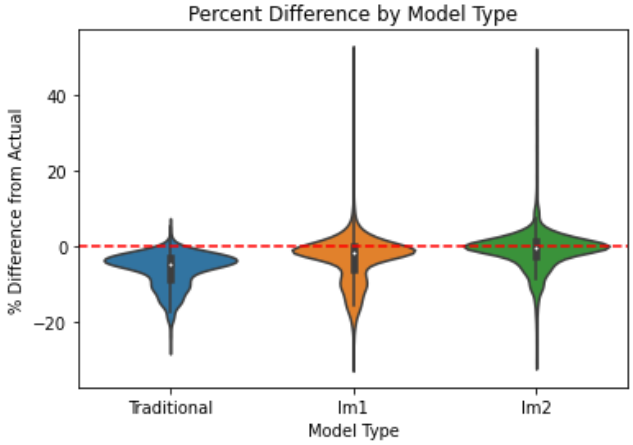

Figure 3: Violin plot summarizing model performance

Overall, it appears the linear models predictions seem to be more centered around the zero point. The annualized model, lm2, was the most densely concentrated around zero. A large portion of the distribution for the rolling linear model, lm1, appears to be concentrated around zero, but more of the distribution lies below zero than the annualized model. Almost all of traditional averaging methods predictions fall below zero indicating it is the least effective in predicting market rents under the market conditions in the dataset used. It is also worth noting both linear models have larger outliers than the traditional averaging method.

To better explore the results, annual means of percent differences were visualized using confidence intervals and violin plots. Note that the rolling and annualized linear models are mathematically identical for the 2016 prediction year since they used the same data points.

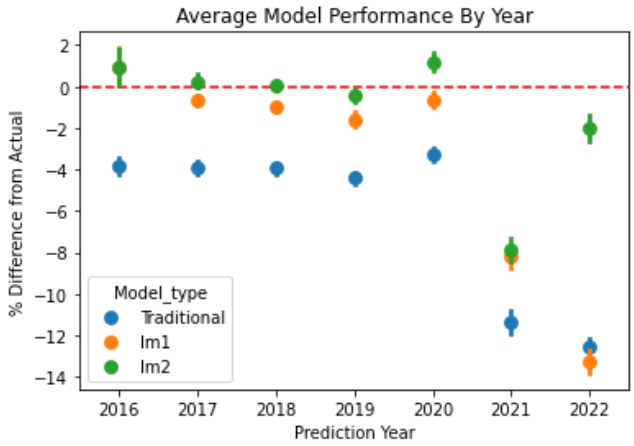

Figure 4: Annual Percent Difference Mean and 95% CI

On an annual basis the mean percent differences were statistically different with the exception of the linear models in 2016 and 2021 as well as the rolling linear model and traditional average in 2022. For all years in the dataset the annualized linear model outperformed the traditional method and generally outperformed the rolling linear model on average. The rolling linear model outperforms the traditional averaging method for all years except 2022. The annualized linear model also significantly outperforms the other models for the most dynamic year in rent changes, 2022. This suggests an annualized linear model might be more favorable in dynamic markets since it is more sensitive to the irregular rental changes.

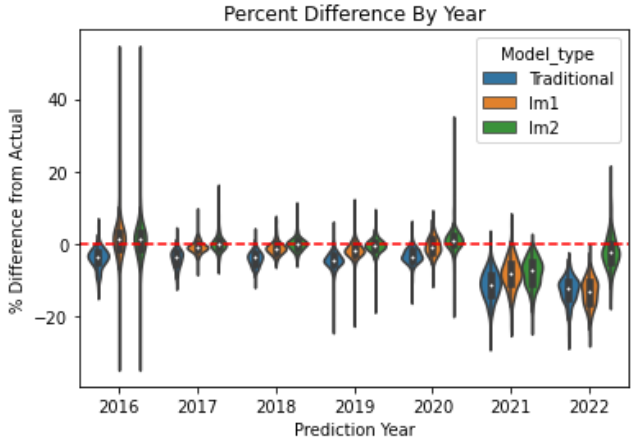

Figure 5:Annual Percent Difference Violin Plot

The annual violin plot highlights that the spread of the distribution is comparable between models within the same year, but the linear models often exhibit larger percent differences outside the interquartile range. It was expected

to see a broader distribution of the linear models in 2016 because there were only 3 data points to build the linear models. It can be seen in subsequent models that the distributions tighten under the normal market conditions and start to loosen up again when rent trends start to deviate in 2021 and 2022.

P-values were used to evaluate if model predictors were statistically significant. Models with a p-value over 0.05 were considered to be statistically insignificant. A summary of statistically insignificant models based on p-values is given in the table below.

Table 3: % of Statistically Insignificant Models by Year

MD\YR	‘16	‘17	‘18	‘19	‘20	‘21	‘22	AVG
LM1	54	10	6	2	1	2	0.5	11
LM2	54	13	8	7	5	19	0.5	15

54% of models were statistically insignificant in the first year iteration. The high amount of statistically insignificant models in the first prediction year was expected due to the few available data points. In subsequent years the percentages varied between 0.5%-13%. In regards to p-value, the rolling linear method appeared to outperform the annual linear method in producing statistically significant models.

To further analyze p-values, the average p-value was plotted by year. Lines represent the 95% confidence interval.

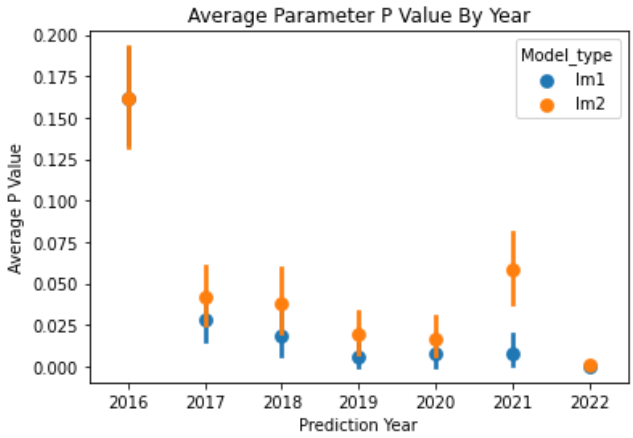

Figure 6: Annual p-value mean and 95% CI

It can be seen from the graph that the rolling model average p-value appears to level out, but the annualized model p-value is more unpredictable. This makes sense given the number of data points increases for the rolling model every iteration, but the number of data points used for the annualized model of each metro area is, at most, 12. This further strengthens that while the annual linear model seems to return the best prediction of rent on average, using date as a predictor alone is not sufficient on an annual basis to reliably predict rent for all metro areas.

R^2 values were used to determine the amount of error explained by the models. It was observed R^2 values tended to be higher for the rolling linear model than the annualized linear model. While the R^2 values do not necessarily indicate how well the models will predict future results, they do suggest that a linear model is a relatively good fit for the data. The plot of the average R^2 values by year is given in figure 7.

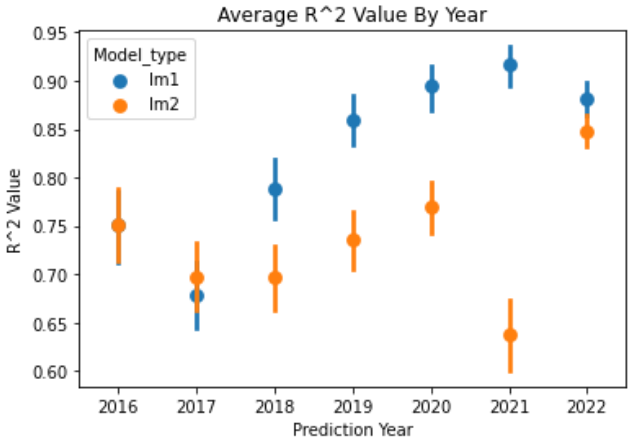

Figure 7: Annual R^2 mean and 95% CI

The dataset was not large enough for efficiency of the models to be a concern, so algorithm efficiency was not evaluated in this study.

5 Discussion

Originally it was thought that the Zillow data might be too sparse to be usable, but this proved to not be an issue. However, it was very difficult to find good data for market specific inflation data. While the original scope of the project involved building models based on inflation and historic rent data, the scope of work had to be modified to only compare the market specific model to the historical averaging method.

Data collection, preprocessing, and warehousing occurred from Oct 31st-Nov 16th, 2022. The preprocessing took longer than expected due to limited internet access while moving to a new state as well as the dataset needing more cleaning/preprocessing than expected. I also had limited experience working with dataframes in Python which

led to a learning curve that slowed down this step in the process.

Data modeling and pattern evaluation was originally scheduled to occur from Nov 7th-14th, 2022 and results were going to be summarized and reported Nov 15th-22nd. However these dates were adjusted due to the previous steps falling behind schedule. Modeling and pattern evaluation occurred from Nov 16th-22nd, and the summarized report was submitted on the 30th of November.

Over the course of the project I was able to learn a couple new libraries and gained a better understanding of working with data frames and linear modeling in Python. I also gained a better understanding of the data mining process and learned to always build more time into the schedule than what you think you initially need.

7 Conclusion

It is important to acknowledge a couple of limitations with the ZORI dataset when interpreting the results. The first is that it is based on data that is mostly available on the internet. It does not capture rental listings that are not advertised online and undisclosed to the Census Bureau. It also does not capture rental increases to tenants that renew leases. Another limitation is related to the overall trends in the data. It appears that most areas experienced rent increases over the covered period so it is unclear how the methodology would work if applied to a dataset that mostly encompassed rent decreases or is relatively flat.

Another consideration is whether the mean rent is actually the best housing metric to build compensation plans if the goal is to project a rate that covers most employees' rent. Given that rent is a distribution, even if one was able to predict rent with 100% accuracy, using the mean rent almost guarantees a large portion of employees will be paying more for rent than what is projected. If an employer wanted to project a rate that ensured their compensation model covered rent for the majority of employees, it might be better to base rates on the 3rd quartile rather than the mean.

In conclusion, using both a rolling linear model and annualized linear model show potential to outperform the traditional averaging method for predicting rent. However, large outliers and some statistically insignificant models suggest more research is required before a model could be adopted for widespread use. In future work, the topics of including a model with home prices as well as breaking down the cost of housing by unit size would be particular topics of interest. Other topics of future work should explore different predictors that may have an influence on rent such as inflation data, population trends, and income trends.

REFERENCES

[1] Shomon Shamsuddin & Colin Campbell (2021): Housing Cost Burden, MaterialHardship, and Well-Being, Housing Policy Debate, DOI: 10.1080/10511482.2021.1882532

[2] Office of Military Compensation Policy. 2022. A Primer on the Basic Allowance for Housing (BAH) for the Uniformed Service. Retrieved on October 30th, 2022 from https://www.defensetravel.dod.mil/Docs/perdiem/BAH-Primer.pdf

[3] Joshua Clark. 2022. Methodology: Zillow Observed Rental Index. Retrieved October 30th, 2022 from https://www.zillow.com/research/methodology-zori-repeat-rent-27092/

[4] The pandas development team. (2022). pandas-dev/pandas: Pandas (v1.5.1). Zenodo. https://doi.org/10.5281/zenodo.7223478

[5] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), 261-272.

[6] Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021

[7] HouseCanary. 2022. National Rental Report. Retrieved 11/28/22 from https://www.housecanary.com/wp-content/uploads/2022/08/HC_2022_Rental-Report_min-5.pdf