IBUYER BUSINESS MODEL ANALYSIS

by

XIAO XIAO

B.S./M.S., University of Colorado Boulder, 2019

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Bachelor of Science/Master of Science

Department of Applied Mathematics

2019

This thesis entitled:
iBuyer Business Model Analysis
written by Xiao Xiao
has been approved for the Department of English

Professor Vanja Dukic

Associate Professor Jem Corcoran

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Xiao, Xiao (B.S./M.S., Department of Applied Mathematics)

iBuyer Business Model Analysis

Thesis directed by Professor Vanja Dukic

This paper focuses on constructing and analyzing different statistical models with respect to an Opendoor dataset from Atlanta during the second half of 2017. Opendoor is one of the iBuyers, an investment company that utilizes technologies along with decades of real estate human-experience to offer homeowners cash for their houses. They typically do minor repairs and maintenance, and then try to quickly re-list the home to sell it for a profit.[1] In this paper we analyzed several regression models including the Simple Linear Regression (SLR), Generalized Linear Model (GLM), and the Generalized Additive Model (GAM), to assess the effects of various house features on profit. The predictors include the preparation days for house listing on market, calendar quarters, zip code, and the square footage for houses. After comparing these models in different situations, we found the GAM with a linear function of square foot and a smoothing function of preparation days produced the best result. Secondly, we changed the response to be qualitative by converting the listed to sold days of houses into binary or binomial based on months. Then, we performed the general Logistic Regression and the GAM logistic model with respect to the binary response and fitted the multinomial logistic regression for the multiple categorical response. Unfortunately, we didn't get ideal results due to lack of observations. However, multinomial logistic regression is definitely a good approach to be discussed in the future with more observations of data. The third section is an introduction to survival analysis, where the attribute of bought to sold days was treated as the survival time and the covariates were square foot, bought prices and quarters. We mainly generated the Cox proportional-hazards model since the Gamma parametric survival model cannot be fulfilled for the three covariates.

Unfortunately, the Cox model didn't present ideal results since the observations are terribly influential individually and some problematic outliers are poorly predicted by the model. Overall, although each model has its own features and advantages/disadvantages, we still need more analysis in the future if a larger set of data is provided so that the models might be improved.

CONTENTS

CHAPTER

TABLES

Table

FIGURES

Figure

Chapter 1

INTRODUCTION

Instant home buyer companies, often referred to as “iBuyers”, have become one of the most popular PropTech (or Real Estate Technology) models over the past few years.[2] Inman Connect has categorized iBuyers as investors that use automated valuation models (AVMs) and other technology to make quick offers on homes, close in days, and then resell them. In fact, Inman[3] recognized iBuyers as the 2017 Inman Person of the Year!

Different from the traditional housing agency, Opendoor, Offerpad, Knock, and Zillow along with others in the revolutionary space of real estate, are referred to as iBuyers. What are the differences? Probably the biggest difference between iBuyers and traditional home flippers is that iBuyers typically don’t purchase distressed properties. Whereas a home flipper might buy the cheapest, most run down house on the block then try to upgrade it to make a profit, iBuyers want homes that are already in decent shape. The iBuyers do make improvements to the homes they purchase, but those improvements are more often cosmetic.[4]

In the traditional way, it may take a long time to search for real estate listings from websites, newspapers, or magazines; as well as to find a reliable real estate agent. Typically, within a few days of the offer being accepted by the house seller, a home inspection is necessary to check for signs of structural damage or things in need of fixing. Moreover, the buyer may also have to work with a mortgage banker to select a loan program, which will then proceed the home appraisement. So, there is a lot of paperwork involved in buying a house, and the waiting time for sellers who may prefer to move quickly is also long.

However, the iBuyer business works in quite a different way, which creates a new option for consumers to make moving much easier. For example, Opendoor operates simply for both sellers and buyers. For sellers, they can directly request an offer to Opendoor which eliminates the hassle of showings and months of uncertainty, so they just need to choose the closing day. After a free home assessment, if repairs are needed, Opendoor can do the work post-close. Sellers can get paid in a matter of days and move to their next home stress-free while Opendoor gets the home list-ready. For buyers, it is much faster to make an offer if they fall in love with a house on the Opendoor website. Further, every home comes with a 30-day satisfaction guarantee so that buyers can purchase with confidence.

Theoretically, iBuyers would probably like sell every home, but in practice they won't buy every house. They actually tend to target middle class homes that were built no later than the mid 20th Century[5] (e.g. for Opendoor, after 1960). Moreover, Opendoor usually focuses on properties that cost between $100,000 and $500,000; while Offerpad typically wants to buy properties built after 1969 and costing between $200,000 and $450,000. [5]

Mike Delprete [6], a scholar at the University of Colorado Boulder and real estate tech strategist as well as an expert advisor, determined the iBuyers median purchase price between November 2017 and October 2018 (see Figure 1), inclusing Opendoor, Offerpad and Zillow. Evidently, Zillow has higher median prices compared with Opendoor and Offerpad.

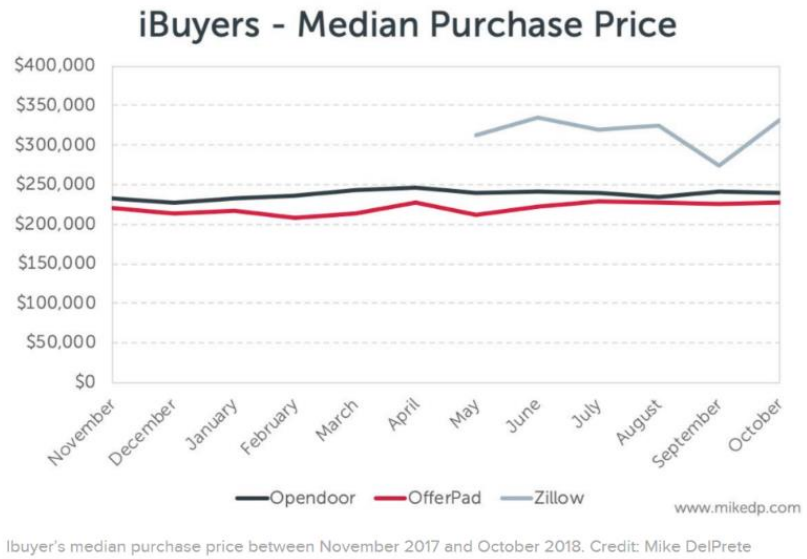

Figure 1: Median purchase price for iBuyers

1.1 Intro to dataset

The project is structured around constructing and analyzing different statistical models, and the main dataset (provided by Dr. Mike Delprete) we used are the houses being bought and sold by Opendoor in Atlanta in the second half of 2017. The important attributes in the original data set include address, bought dates and purchase prices, listing dates and prices, sold dates and prices, and the profit (the difference between bought prices and listing prices).

Since there were some missing values with regard to the sold dates and prices, we searched and completed the houses' sale information from the Zillow website (https://www.zillow.com). We also added a new column of the corresponding "zip code", because it was not included in the address. In addition, it's apparent that the profit differs due to the floor sizes of the houses. In order to be consistent, we also

added a new attribute of “square.footage” so that we can compute the profit per square foot for better comparison.

1.2 Purpose

How does Opendoor make money? There are two ways: 1. from the service fee it charges, 2. from any difference between purchase prince and sale price.[6] Therefore, studying the profit always comes first in statistical analysis.

The project firstly focused on analyzing and predicting different effects on profit per square foot which is called the “dependent variable” in statistics. The predictors (also known as “explanatory variables”) were considered first to be as follows:

- Zip code
- Square footage of each house
- Preparation days (from bought to listed dates)
- Season/quarters when the houses are listed for sale: Q1 (January - March), Q2 (April - June), Q3 (July - September), and Q4 (October - December).

In this situation, we began with the simple linear regression models to find a better combination of predictors. For example, the “quarter” cannot be used. It was not significant in the models because we don't have any houses bought or sold in Q1 or Q2. It is not strong enough to predict the profit per square foot. In the modeling, we also tried log-transformation and square root transformation of the preparation days manually because of the visual non-linear relationship between the days and profit. Furthermore, the Generalized Linear Model with a Gamma regression was also performed as an extension of the simple linear model, where the response

variable was not restricted to be Normally distributed. Next, we applied the Generalized Additive Model to allow for non-linear relationships between response and predictors by using smoothing functions automatically.

Secondly, we treated the "listed to sold days" (the period from when Opendoor listed the house back on market to the date when the house was sold again) as a qualitative variable instead quantitative in order to see whether or not the floor size of houses and purchase prices have strong relationships with the listed to sold days. In this case, the original quantitative variable could be converted into binary variables 0/1, and we applied the logistic regression. That is to say, for example, the new response would be equal to 1 if the listed to sold days are within a month and 0 if not. Furthermore, if classifying the listed to sold days into 4 levels of dummy variables separated by month (e.g. level 1 is within 1 month, level 2 is between the first and second months, and so on), we would obtain a multiple (more than 2) categorical response variable, which could be discussed by using the multinomial logistic regression model.

The next section is the survival analysis with a semiparametric method called Cox proportional hazards regression model and a Gamma parametric survival model, in order to measure the risk of covariates (including square foot, bought prices and quarter) to the bought to sold days (the date of buying house to the date of selling). For this model, we used the original set of data without replenishing the information from Zillow, because one should care about the censoring of observations, which will be discussed later.

Chapter 2

LINEAR AND NON-LINEAR REGRESSION

First of all, we created a column of variables called "Quarter" which correspond to the bought dates of house. Then, we calculated the profit per square foot (denoted by "prof.sqft"), which was regarded as the response or dependent variable.

Before modeling, we made a matrix of plots through the function *ggpairs* in R (using the package "GGally"), including scatter plots, box plots, histograms and density plots, as well as the correlation coefficients. The purpose is to see the correlation of 4 important variables: profit per square foot (sqft, for short), sqft, quarter, and preparation days. Referring to Figure 2, to decide rough relationship between "prof.sqft" and the other three variables, it's obvious that there is a negative correlation between "sqft" and profit per sqft. Since we only have the data for the houses during Q3 and Q4 (which is to say, from July to December) and we only have 20 home sales available for analysis, there is not enough evidence to say that winter could bring more profit than autumn, although the boxplot shows higher profit in Q4 than in Q3. In addition, the preparation days seem to have only a small impact on the profit per sqft, since the correlation coefficient is only 0.152. The number of preparation days are generally around 15 in total from the density plot. Moreover, we should also notice that the correlation between sqft and preparation days is relatively strong and positive. To be specific, matching real life expectations, the larger the home is, the greater the number of preparation days for fixing and maintenance.

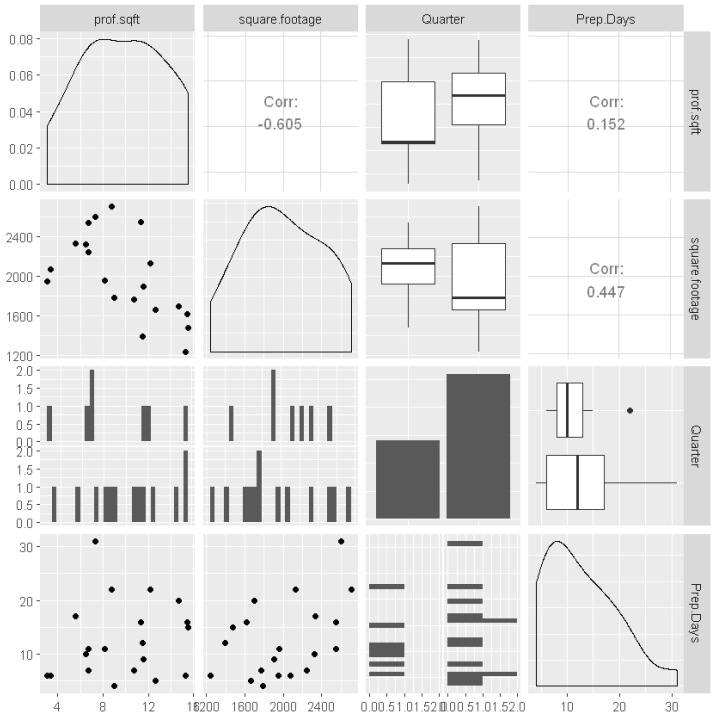

Figure 2: Correlation of variables

2.1 Simple Linear Model

To begin, we fitted the most basic simple linear regression model with 4 predictors: square footage, preparation days (from bought to listed dates), bought to sold days, and quarters. Unfortunately, since the dataset only contains 20 houses in Atlanta, and all homes were sold within Q3 and Q4, it's not statistically significant to discuss the relationship between quarters and profit per square foot. The model was also fitted meaningless if using the interaction terms such as "Quarter × Preparation days". In addition, from the summary of this model, we found that the variable of bought to sold days doesn't need to be considered, since the p-value (equal to 0.2105) shows less significance to describe its effects on the profit.

Therefore, after reducing the insignificant variables, the only predictors we cared about in this model should be "square.foot" and "preparation days". As for the model fitted in *R*, the R^2 value was 0.5895 (this is also the proportion of the variability in profit per sqft which is explained by the two predictors), and the Akaike Information Criterion (AIC) value we got was 100.34. Moreover, from the coefficients of the model, as shown in Table 1, we should notice that the profit per square footage will decrease by 0.77 if the area of house increases by 100 square feet. Inversely, it will increase by 0.29 for an additional day of preparation before resale. However, since the values of square footage and preparation days can never be equal to 0, the intercept has no intrinsic meaning in this situation.

Table 1: Coefficients of SLR

(Intercept)	**square.footage**	**Prep.Days**
21.548579	-0.007703	0.288336

To check how well the simple linear regression (SLR) model fits or summarizes the data, we refer to Figure 3 below. It shows four diagnostic plots: "Residuals versus Fitted Values", "Normal Probability Plot of Standardized residuals" (or "Q-Q plot"), "Square root of Standardized residuals versus Fitted Values", and "Leverage-Residual and Cook's Distance". Roughly speaking, the SLR model's fit is not bad. However, to be specific, we need to check whether or not the four types of assumptions are valid in this case by considering Residuals, Predictors, Model form, and Observations.

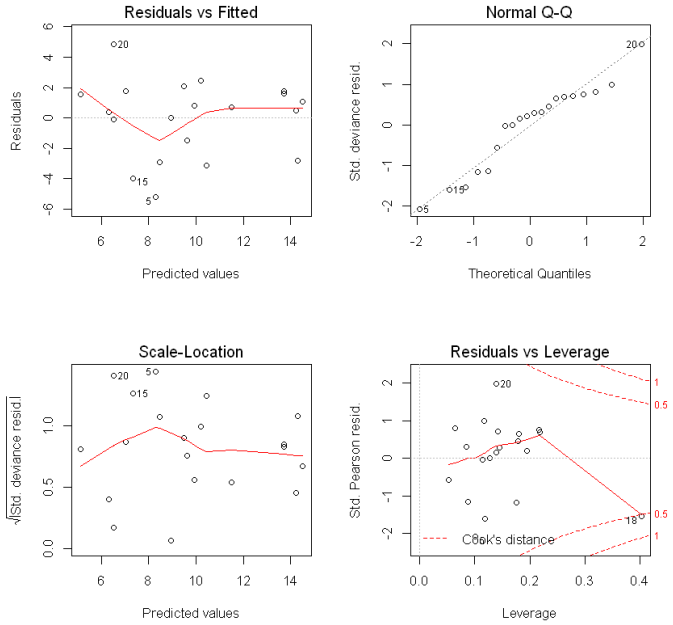

Figure 3: Diagnostic plots for simple linear model

Residuals: Checking the Residual against the Fitted value plot, we can see that the residuals are mostly distributed randomly around zero, although there is a little curvature which shows some non-linear patterns. In consideration of the non-random pattern, we used the ACF ("Autocorrelation function") plot to show the correlation of residuals (as a time series) with its own lags (or lag operator), as Figure 4. The horizontal dashed lines represents lag-wise 95% confidence intervals centered at zero and indicate bounds for statistical significance. Overall, most of the correlations are small enough within the confidence intervals, which implies the no-autocorrelation assumption. However, some correlations at lags 2 and 5 are not particularly small, which are probably caused by contingent factors because the dataset is not large enough. Therefore, we can conclude the following:

- linearity is not violated
- errors have constant variance

- errors appear uncorrelated.

In addition, Multiple regression assumes that the residuals are Normally distributed, in this case, we can use the Q-Q plot to assess it; specifically, the standardized residuals exhibit close to a Normal distribution if the observed quantiles are all fairly close to the "$y=x$" line. Based on the plot we got, the residuals do not follow the straight line perfectly, but are not bad. And there are two notable outliers, with standardized residuals around 2 and -2: observations number 5 and 20. Generally speaking, the standardized residuals are just close to the Normal distribution, but we need to more data to confirm it in the future.

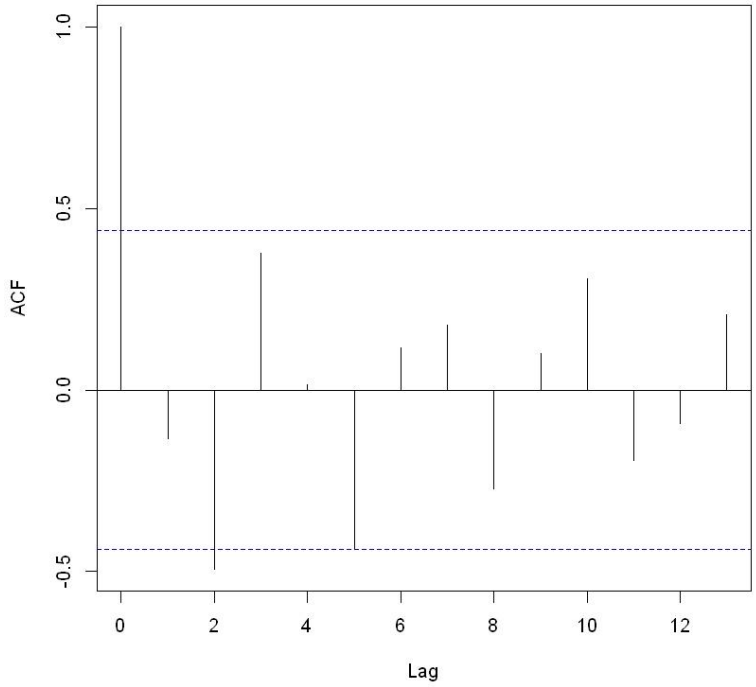

Figure 4: ACF for residuals

Predictors: The only assumption that can be checked is linear independence, which could be very roughly checked by the correlation matrix (see Figure 2) or a

correlation matrix (see Table 2) since examining pairwise correlations does not tell us much about larger linear combinations. Having converted the "Quarter" into numbers ("Q3" to "3", and "Q4" to "4") only, we obtained the correlation values in Table 2. Clearly, "prof.sqft" and "square.footage" are somewhat correlated, which is expected since the "prof.sqft" is directly calculated by the *sqft* unit. But the other two predictors, "Quarter" and "Prep.Days", do not show very strong correlation.

Table 2: Correlation matrix

	prof.sqft	**square.footage**	**Quarter**	**Prep.Days**
prof.sqft	1.000	-0.605	0.177	0.152
square.footage	-0.605	1.000	-0.147	0.447
Quarter	0.177	-0.147	1.000	0.129
Prep.Days	0.152	0.447	0.129	1.000

Model Form: The residual plots indicate no major linearity violations in the simple linear regression model. Furthermore, in order to model the residuals of one predictor against the dependent variable and see the linear relationship, we displayed two Component Residual plots ("crPlots" in *R* through the *car* package), which are also known as an extension of partial residual plots. This is shown in Figure 5 with the dashed line indicating where the best fitting model lies. First, with respect to the direction of dashed lines, we notice that the profit per square foot will decrease as the area of home increases, if one holds the preparation days to be constant. Conversely, the profit per sqft will increase for longer preparation days. Next, the green curve implies there is no significant difference between the residual line and the component line for the first predictor, "square.footage", which indicates that it has a linear relationship with the profit per sqft. However, the second predictor,

"Prep.Days", shows some insignificant patterns, although the problems are not noticeable. There are some methods to "correct" these differences, such as changing the form of predictors. Typical alterations are *sqrt*(Prep.days) or *log*(Prep.days) in this situation due to the shape of the residual curve. Modeling again using the two alterations, we found the percentage of the variation in "prof.sqft" that can be accounted for by the two predictors (refer to the multiple R^2 value) has increased by 2% in two ways. The new component residual plots with a transformation of *log*(Prep.Days) are shown in Figure 6. Although it's not a big change, the *log* form of preparation days shows better and more significant relationship to the dependent variable.

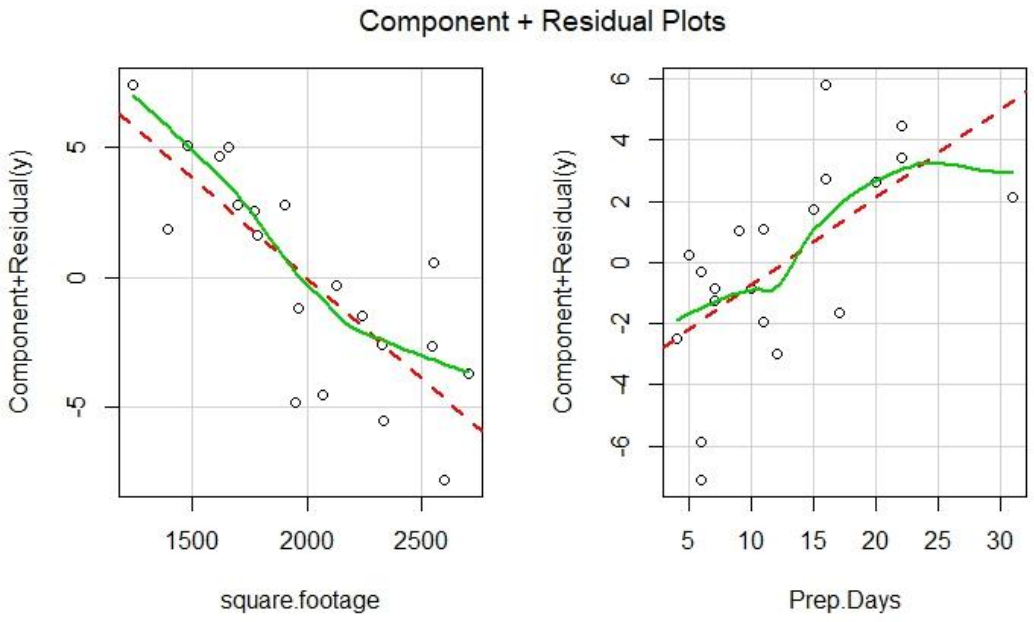

Figure 5: Original component residual plots

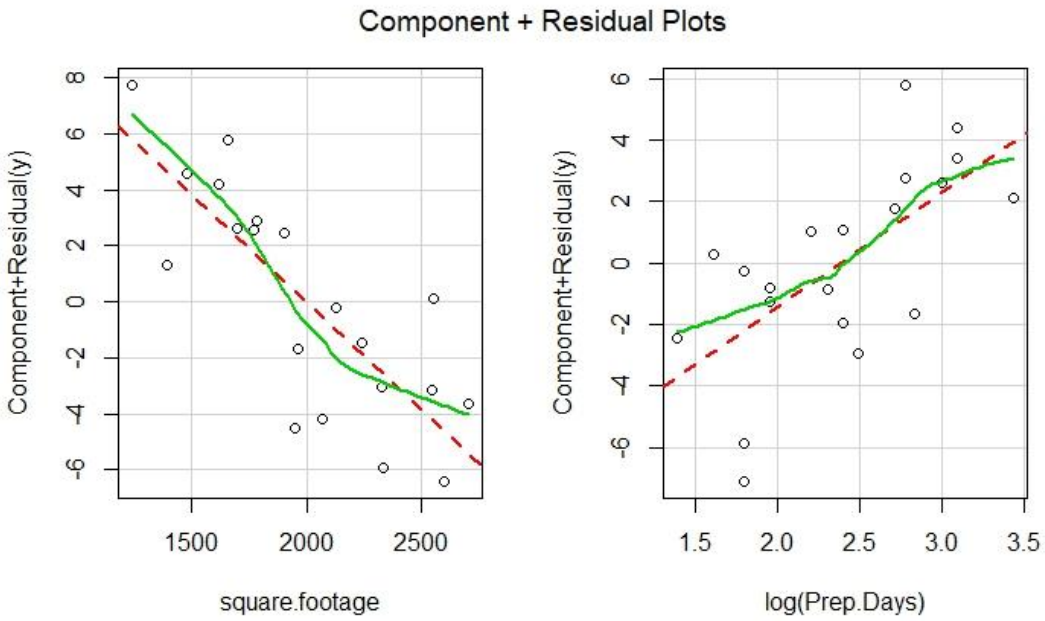

Figure 6: CR plots with *log*(Prep.Days)

Observations: The main assumption regarding the observations is that points have approximately equal influence on the fitted response values. Referring to Figure 3, either the Cook's distance or the Residuals versus Leverage graph can be used to examine the influence of and identify outliers (observations for predictor variables). From the Cook's distance graph, observation 18 has Cook's D over twice as large as any other, and observation 20 is also influential, though to lesser degree. The leverage-residual plot gives similar results: number 18 had giant leverage, while number 20 has a combination of high leverage and high residual. Therefore, these two points need to be examined further. Observation 18 only has 4 preparation days, while the average value for our dataset is 12.65 (close to two weeks). Observation 20 has a small house area but gains a relatively large amount of profit per square footage. This is somewhat suspicious and not representative, so we still need advanced investigation and perform more complicated model in further study.

2.2 Gamma Regression as Generalized Linear Model

Ordinary linear regression predicts the expected value of a given unknown quantity (the "response variable", a random variable) as a linear combination of a set of observed values ("predictors").[11] This implies that a constant change in a predictor leads to a constant change in the response variable (i.e. a linear-response model).[11] In addition, since the error term has a Normal distribution, the ordinary linear regression model should be appropriate when the response variable has a Normal distribution. However, in certain cases, the response variable cannot be or is only approximately Normally distributed. In these cases, one can try to identify a transformation for the continuous response variable (denoted by y), e.g. $y' = log(y)$ or $y' = \sqrt{y}$ are typical, in order to obtain new data, y', which is closer to Normal.

In this case, to develop and extend the simple linear regression model, we could consider and apply the generalized linear model (GLM) by Nelder and Wedderburn (1972) in the next step, which is a flexible generalization of ordinary linear regression which allows for response variables that have a distribution other than a Normal distribution [11] (they may even be categorical rather than continuous variables). Nevertheless, there is no guarantee that the response and explanatory variables must have a simple linear relationship. To be more specific, a generalized linear model consists of three components as below, referring to Dobson [13] (2002).

- The set of response variables $Y_1, \ldots, Y_n$ are assumed to share the same distribution from an exponential family, such as Poisson, Gamma, Binomial.
- A linear predictor $\eta = \boldsymbol{X}\beta$, where β is a set of parameters, and $\boldsymbol{X}$ is the design matrix of explanatory variables.
- The linear model is related to the response variable via a monotone "link function" g such that $E(Y_i) = \mu = g^{-1}(\eta)$.

Therefore, based on the knowledge of GLM above, we could firstly see whether or not the profit per square footage is Normally distributed by looking at the histogram with a density curve as shown in Figure 7. Unfortunately, the density curve doesn't show a strong Normal distribution, which has a large variance value. However, if we regard the response variables Y_i as being approximately Normally distributed, we could use a Gaussian in GLM in R with the canonical link function as the identity function, since $g(\mu_i) = \mu_i = E(Y_i)$ and $Y_i \sim N(\mu_i, \sigma^2)$. This situation should be identical to the simple linear regression model in section *A*.

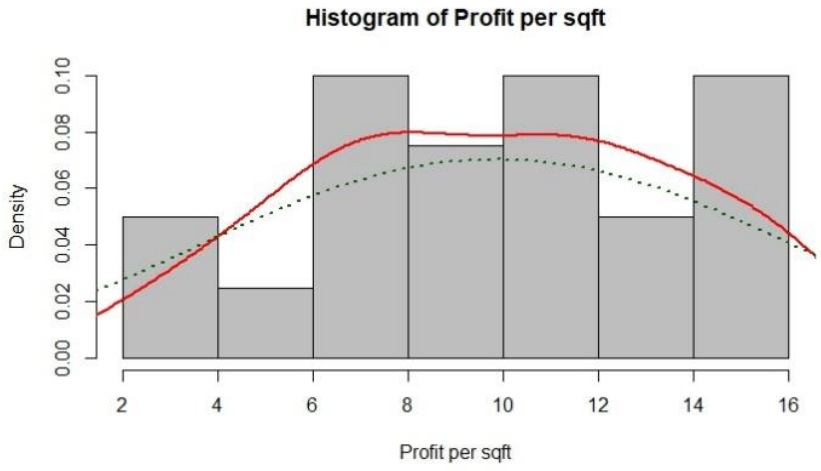

Figure 7: Histogram of Profit per sqft

On the other hand, referring to the smoothed curve (dashed green line) in Figure 7, since the histogram is not symmetric around the mean and the distribution of profit per sqft is slightly skewed left (also known as negatively skewed) and always positive, the most common method of analysis is to perform the simple linear model with log-transformed outcome. Unfortunately, the result is even worse than the model discussed in section *A*. However, an alternative way is to apply the GLM with

Gamma regression, while the link function can be determined by comparing the canonical “inverse” link and the “log” link function which is consistent with the log-transformation. Although there is not much difference between these two link functions, the Gamma regression with “log” link is more adequate a model than the other one.

Therefore, we need to compare the two generalized linear models with Gamma and Gaussian distributions by noticing the differences in two ways. First, from the Residuals versus Fitted value plots of these two models, as shown in Figure 8, it is apparent that the Gamma regression with log link fits better than the Normal distribution with identity link, due to the randomness of residuals. Second, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are commonly used estimators for model selection that cannot be ignored. They provide a method for assessing the quality of a model through comparison of related models for a given data set. And the AIC and BIC in isolation are meaningless. Rather, for different candidate models, the model with the smallest value of AIC and BIC is preferred. Although AIC and BIC cannot always be used together for every case and model, we don't need to care about the difference, since our data set is small enough, finite, and the number of predictors is limited. Based on these statistics, the GLM with Normal distribution should be chosen since the AIC and BIC are slightly smaller as shown in Table 3, which is inconsistent to the conclusion from the former comparison.

Table 3: Gamma and Gaussian comparison

	AIC	**BIC**
Gamma	106.85	110.84
Gaussian	100.34	104.33

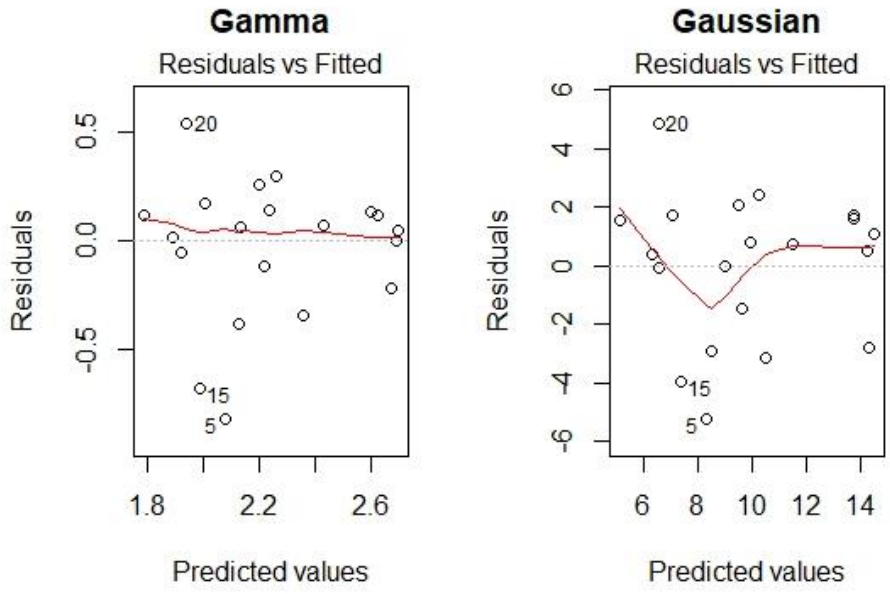

Figure 8: Residuals vs. Fitted plots for GLM

Although the AIC and BIC estimators are informative, we prefer not to use the conclusion of them since the values are quite close for the two models. Overall, the Gamma distribution with "log" link fits better than the Normal distribution with "identity" link function. Furthermore, to test how the Normal distribution and Gamma regression as GLM fit the actual values of profit per square foot, we plotted the prediction line (see Figure 9) for both of the models. Unfortunately, it is not obvious to see any difference between the prediction lines and conclude which model fits better. In this case, these two models cannot be considered the same, since the observed values are extremely limited to display any problematic pattern and discrepancy. Therefore, we still need larger set of data in the future for further study.

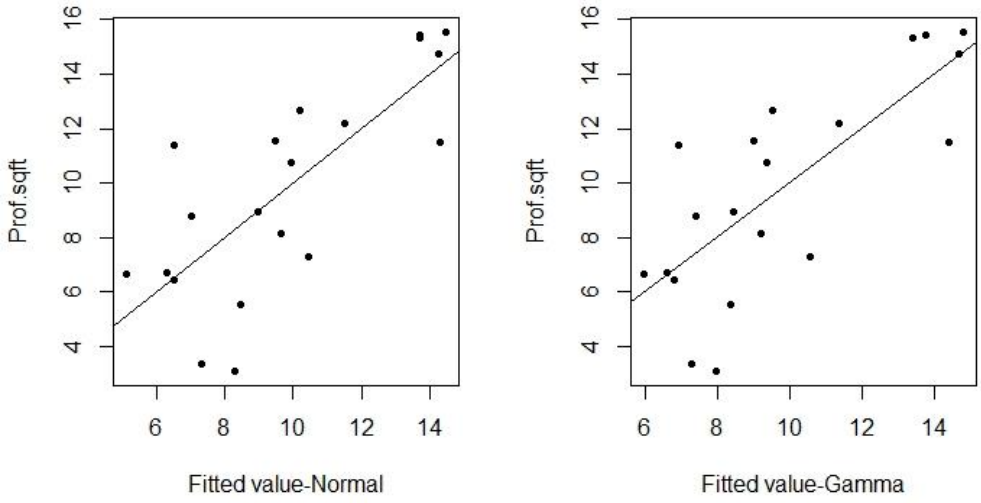

Figure 9: Prediction for Normal and Gamma

2.3 Non-linear Modeling: Generalized Additive Model

Since the set of data we have is extremely small and limited, even if the Gamma regression as generalized linear model performed well, we cannot guarantee that the model in the presence of more data should be linear. In order to relax the assumption of linearity between predictor variables and response variable, it's a typical method to use the Generalized Additive Model [14] (GAM, originally created by Hastie and Tibshirani, 1990) which is another extension to the Generalized Linear Model (GLM) performed in the last section, where the linear predictor η is not restricted to be linear in the covariates, **X**, but is the sum of smoothing functions applied to the x_i's [11], while maintaining additivity.

To be detailed, the regular multiple linear regression model has the linear predictor form

$$\eta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}.$$

However, to allow for non-linear relationships between each feature and the response [15] (see G. James et al., 2013), the linear components are replaced by the following expression, and η becomes an additive predictor.

$$\eta = g(E[Y]) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n),$$

where β_0 is the intercept, x_i are the predictor variables, Y is the response variable, and f_i are smooth functions which are unknown and estimated from the data. It is called an *additive* model because we calculate a separate f_i for each X_i, and then add together all of their contributions.[15]

Back to our example, the model takes the form

$$\text{profit.sqft} = \beta_0 + f_1(sqft) + f_2(prep.days) + f_3(quarter) + \varepsilon,$$

where ε is the error term, and "sqft" stands for square footage, then "prep.days" is corresponding to the preparation days. Although "quarter" has been included in the above model for completeness, it can never be significant in each model for the given data set, so it is again discarded in the GAM analysis.

In order to be consistent with the simple linear regression model in section *A*, as an example, we could fit a GAM to predict the profit per square foot by applying natural spline functions of "sqft" and "prep.days". This is accomplished in *R* with "*ns*(sqft)", and the model can simply be displayed by l*m*() function. Next, smoothing splines should be used instead of natural splines. To perform the GAM in *R*, we will need to install two packages: "gam" and "mgcv". The regression would change to use "*s*(sqft)" function which is part of the *gam* library. The *s*() function is applied to indicate the use of a smoothing spline.

Initially, with both of the terms (sqft and prep.days) fitted simultaneously using the smoothing function, we got two plots as shown in Figure 10. The left-hand

panel implies that holding preparation days fixed, the profit per sqft tends to decrease fast with the field area of houses. However, the right-hand panel indicates that holding the square footage fixed, the profit per sqft tends to be increase at a larger rate when the preparation days are smaller than 20, while increasing slower after 20 days.

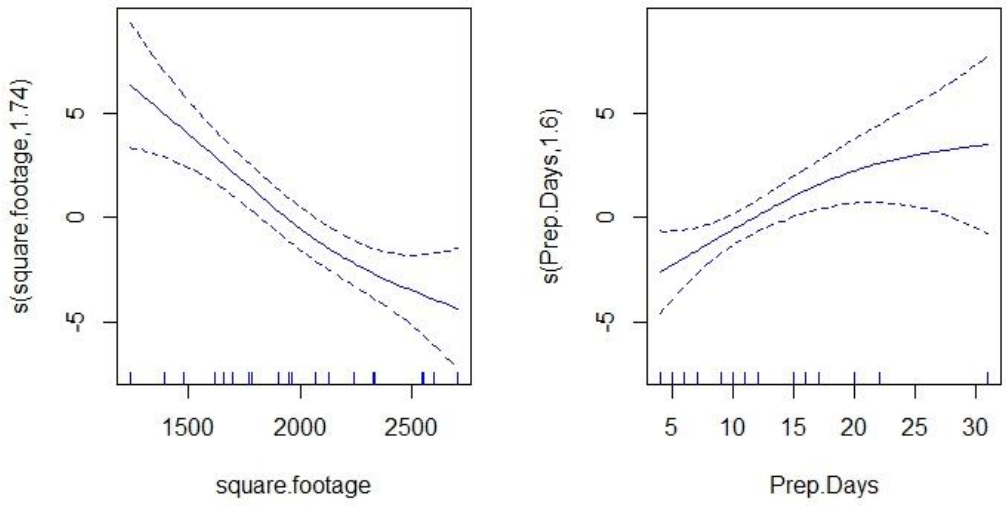

Figure 10: Plots of Generalized Additive Model

Nevertheless, since the function of "sqft" looks rather linear, we first tried to remove this predictor and keep only the preparation days with a smoothing function; alternatively, we discarded the smoothing function of sqft and left it as is while keeping *s*(prep.days). This allowed us to perform a series of ANOVA tests in order to determine which of these three models is preferred: a GAM that excludes sqft (called M_1), a GAM that uses a linear function of sqft (called M_2), and a GAM that uses a spline function of sqft (called M_3). The analysis of deviance table from ANOVA tests is as Figure 11.

Model 1: y ~ s(Prep.Days) Model 2: y ~ square.footage + s(Prep.Days) Model 3: y ~ s(square.footage) + s(Prep.Days)						
	Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
M_1	17.318	267.229				
M_2	16.189	109.851	1.1286	157.38	22.565	0.0001857 ***
M_3	14.858	96.798	1.3308	13.053	1.5871	0.2337558

Figure 11: Generalized Additive Models Comparison

We find that there is compelling evidence that a GAM which uses a linear function of square footage and spline function of preparation days is better than a GAM that uses spline functions for both of the predictors, since the p value (= 0.0001857) is much smaller. However, there is no eveidence to say that a non-linear function of square footage is needed (p-value = 0.234). In other words, M_2 is the best model among these three GAMs, regarding to the outcomes of the ANOVA test.

In general, the GAMs seem to be more resonable than the linear regression models, however, the advantages ("●") and disadvantages ("○") need to be discussed as the following aspects.

- The GAMs are not limited to linearity and allow us to fit non-linear smoothing functions $f_i(x_i)$ to each corresponding predictor X_i. As a consequence, one is able to simply and automatically posit non-linear relationships between the predictors and response variable which linear regression models cannot accomplish. In this way, if any non-linear relationships exist, the GAMs will be a better and more accurate approach to predict the profit per sqft than transforming the predictors by *log* or *sqrt* functions by hand as we tried in the simple linear regression models.

- Since the GAMs are additive, it is much more efficient to analyze the influence of each individual explanatory variable on the response with other the variables remaining unchanged than performing the added-variable plots (partial regression plots) and component-residual plots (partial residual plots) among the linear regression models.
- A main limitation of GAMs in our study is the propensity to overfit. Referring to the comparison of GAMs discussed by the ANOVA test, the model that has spline functions applied to all the predictors is not optimized as we expected, while the original linear function for "square.footage" works better.

Chapter 3

LOGISTIC REGRESSION

The linear regression model analyzed above assumes that the response variable "profit per sqft" is quantitative, and the non-linear GAMs are also examined for the quantitative response, although one can apply qualitative response variables to GAMs as well. Alternatively, the common way is to fit logistic regression models to the qualitative response, which would be binary, is as follows:

$$Y = \begin{cases} 0 & if\ A\ happens; \\ 1 & if\ B\ happens\ OR\ A\ doesn't\ happen; \end{cases}$$

where A and B are two possible events to describe the response variable, such as yes/no, pass/fail, win/lose, or alive/dead [13]. For example, the committee working in the department of Applied Mathematics in the University of Colorado Boulder must be able to determine whether or not an applicant who applied the graduate program can be admitted as a good candidate, on the basis of the student's GPA, GRE scores, personal statement, resume evaluation, and so forth. In this case, we could use the indicator variable to imply the binary response as *Y=1* if the applicant is admitted, while *Y=0* if not.

In addition, logistic regression is also known as a Generalized Linear Model with binomial distribution for the response Y and the link function "$\eta = logit(\pi)$", where π is the probability that the binary response *Y=1* for a given set of data. That is to say, π is the probability that a applicant is admitted into the graduate school in the previous example. Consequently, the general logistic regression model is as follows:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x_i^T \beta}$$

where $\boldsymbol{x_i}$'s are the predictors and $\boldsymbol{\beta}$ is the parameter vector.

3.1 The Logistic Model

In our project, to explore a situation when the response is instead qualitative, we could classify the preparation days (from bought to listed house dates) or the sale days (from listed to sold dates) into several groups, which will be referred to as categorical. However, since all of the preparation days from bought to listed dates from Opendoor are within a month (the longest preparation is 31 days in the data), classifying the listed to sold dates becomes more appropriate.

To decide how to classify the new response variable "listed.sold days", we would like to firstly see its histogram as Figure 12, from which we noticed that most of houses were sold within about 30-40 days from the listing dates by Opendoor. By calculation, the minimum listed to sold days is 24 with respect to our data set, while the maximum days is 150 which is equivalent to 5 months. So, since the range (greater than 4 months) of days is large enough, we determined to classify them into 4 groups: less than 30 days, less than 60 days, less than 90 days, and greater than 90 days.

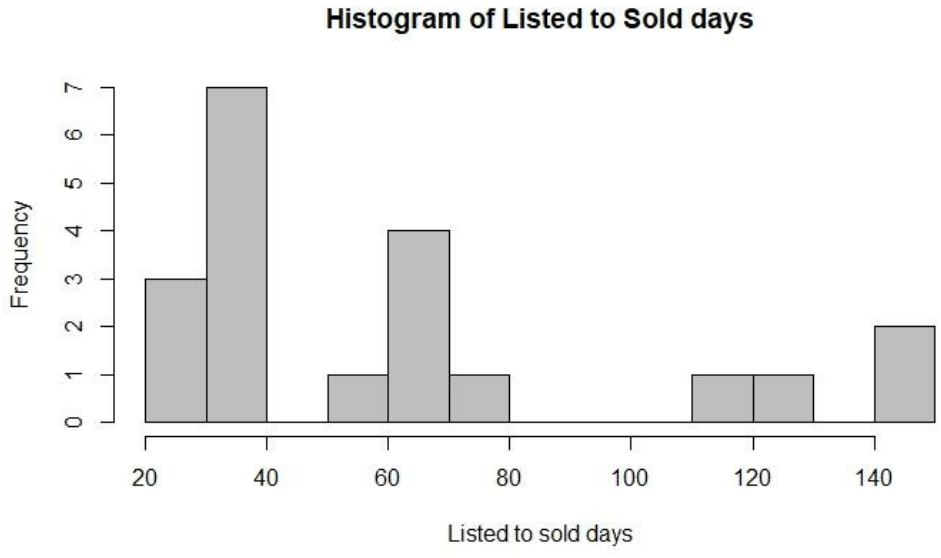

Figure 12: Histogram of Listed to Sold days

In this way, there would be 4 logistic regression models that could be generated, where the response *Y* should be. For example, *Y=1* if the listed to sold days are less than 30, while *Y=0* if not (which means the days are greater than 30) as the first model. The form would be as follows in this case:

$$log\left(\frac{P(Days < 30)}{P(Days > 30)}\right) = \beta_0 + \beta_1(sqft) + \beta_2(bought.for)$$

where "bought.for" is the purchase price that Opendoor paid the customer who sold the house.

To run the logistic models in *R*, we could use the GLM model with the family of "binomial" as well as the "logit" link function. The diagnostic plots for residual versus fitted values are shown as Figure 13. It's obvious from the plots of the first two groups that listed to sold days less than 30 or 60 show some problematic patterns, which may be cause by the lack of data.

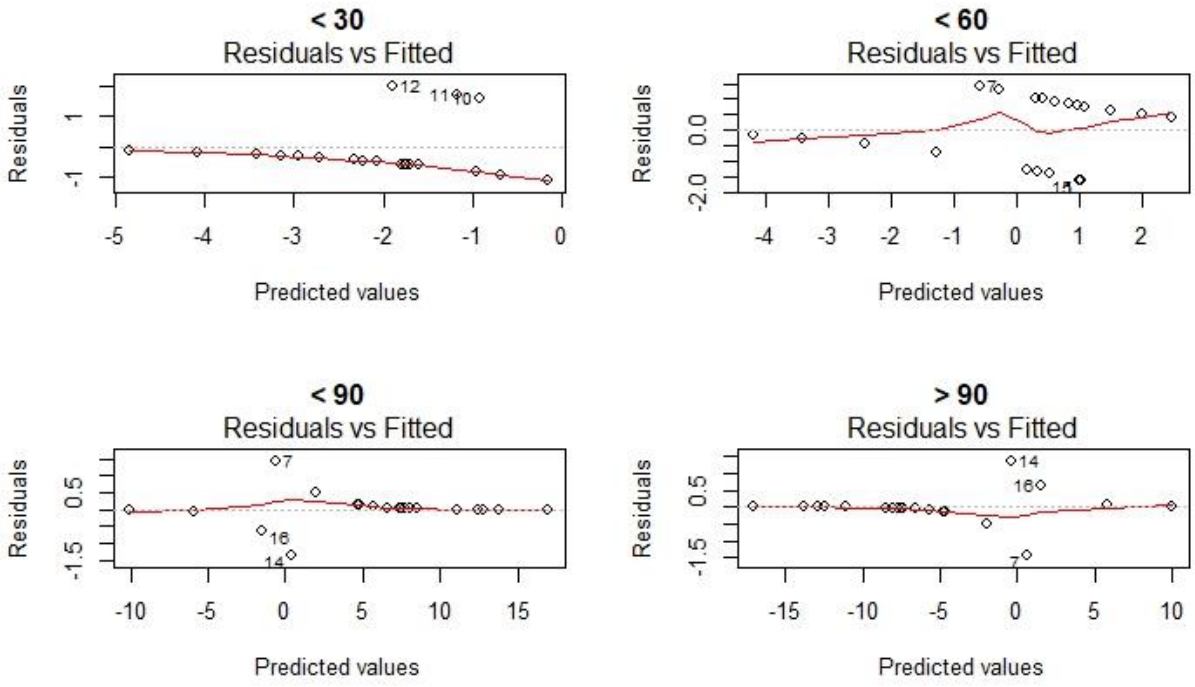

Figure 13: Residual vs. Fitted of logistic models

However, the other two groups, listed to sold days less or greater than 90, have fitted much better with more evidence of data. Therefore, as an example, we'd like to take a look at the summary of the logistic regression model when the days are less than 90, as shown in Figure 14. This summary can be discussed in terms of the coefficients, p-values, and AIC.

```
glm(formula = I(Listed.Sold.Days < 90) ~ square.footage + Bought.For,
    family = binomial, data = atlanta)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.36616   0.00111   0.02271   0.06009   1.44211

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.769e+01  3.249e+01   1.160    0.246
square.footage -9.086e-03  8.567e-03  -1.060    0.289
Bought.For     -6.172e-05  4.888e-05  -1.263    0.207

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20.0161  on 19  degrees of freedom
Residual deviance:  4.6452  on 17  degrees of freedom
AIC: 10.645

Number of Fisher Scoring iterations: 9
```

Figure 14: Logistic regression (days less than 90)

- **Coefficients.** The coefficients show the relationship between square foot, bought prices, and the listed to sold days which is on the logit scale. Both of the predictors have negative coefficients, which suggests that if the floor size of house and the purchase price increase, we would get a decrease in the log-odds of the response. It means the probability of house being sold within 90 days will decrease and the period of house sale will tend to be longer. These facts are also applicable to the other two models when listed to sold days are less than 30 or 60; while it is inverse when the days are larger than 90.

- **P-values.** We notice that none of the p-values are significant in the summary, where the smallest one is associated with bought prices. Hence, since the p-values are relatively large, there is no strong evidence of real association between the predictors and response.
- **AIC.** The Akaike Information Criterion (AIC) value is pretty small in this case. However, AIC in isolation is not meaningful and is relative to each of the other models. So, we'll discuss it later for models comparison.

Next, as mentioned in the beginning of logistic regression, GAMs can also be applied in the situation when Y is qualitative. The logistic regression GAM allows for non-linear relationships by taking the following form:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + f_1(x_1) + \cdots + f_n(x_n)$$

where f_i is the smoothing function. Based on the formula, we fitted the logistic GAM in R with the model

$$log\left(\frac{P(Days < 90)}{P(Days > 90)}\right) = \beta_0 + f_1(sqft) + f_2(bought.for).$$

Then, the resulting fit of this model is shown in Figure 15, which shows basically consistent relationship to the general logistic regression. However, the square footage shows more linearity, while the bought price has non-linear relationship with listed to sold days. Both of the predictors still maintain the negative relationships to response. Although, the p-values of predictors still indicate insignificance, the AIC value (= 7.71) is slightly smaller than the general logistic regression model.

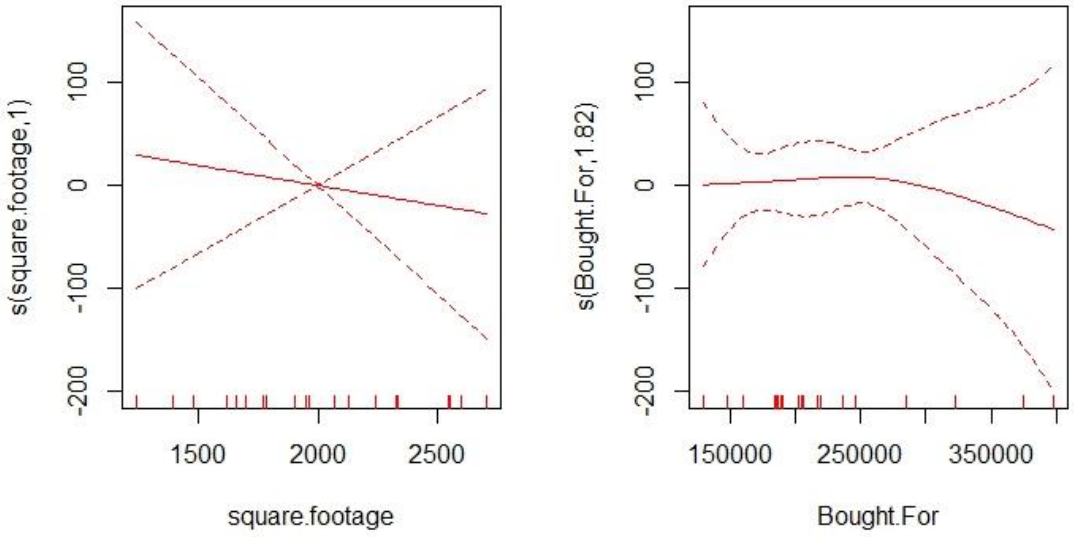

Figure 15: GAM – Logistic regression

3.2 Multinomial Logistic Regression

Before next step, we would firstly like to separate the full data set into two subsets: one is when the listed to sold days (LSD) are less than 90; and the other is when LSD are larger than 90. Then, we plotted 4 histograms in order to look at the distributions of square foot and bought prices with respect to the two subsets. As shown in Figure 16, the blue vertical lines marked the mean values of two histograms. This indicates that Opendoor often needs more time until the houses are sold again if the floor sizes are larger since the mean is around 2400 sqft when LSD are greater than 90 while the mean is only about 1900 sqft when LSD are smaller. As for Figure 17, the difference between two histograms implies that the house with higher bought

price often need more time to be sold, which is consistent with the phenomenon in Figure 16.

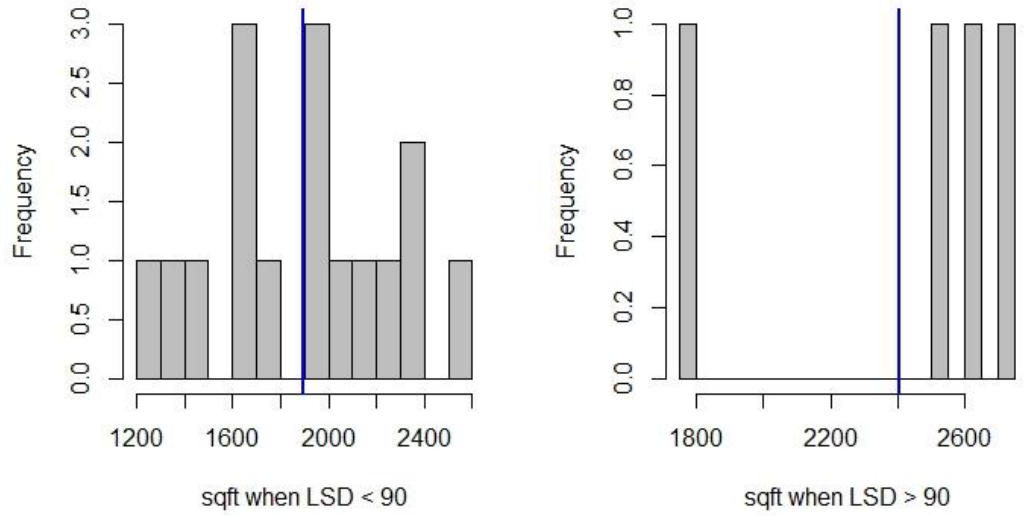

Figure 16: Histogram of SQFT regarding to LSD

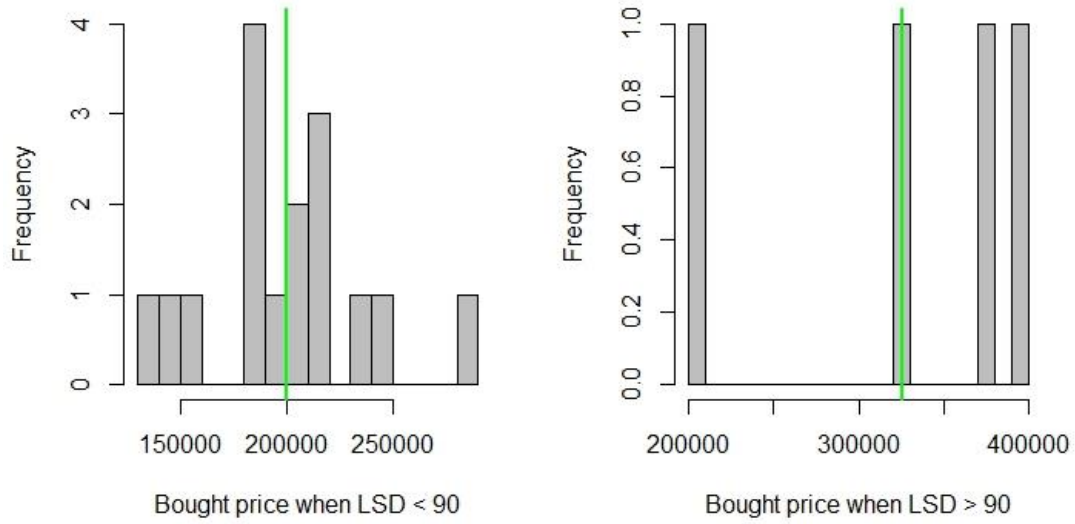

Figure 17: Histogram of Bought Price regarding to LSD

Based on the 4 logistic regression models in last section, we were curious about a logistic model that could describe and predict the 4 categories together, which can be achieved by the Multinomial Logistic Regression (MLR). And MLR is an extension of logistic regression used when the categorical response variable (unordered) has more than two categories and to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable given a set of independent variables (see from *Wikipedia*).

Therefore, to fit MLR, we should firstly create a dummy variable based on the 4 possible outcomes of listed to sold days, which is, to define "LSD < 30 == 1", "30 < LSD < 60 == 2", "60 < LSD < 90 == 3", and "LSD > 90 == 4". With these categories, we'd like to output some box plots of the square foot, bought prices (in units of thousands of dollars) and preparation days with respect to each group as displayed in Figure 18. The bold black lines represent the median values in each variable. And the common feature is that the larger the floor sizes, bought prices, and preparation days are, the longer period the houses need to be sold. Next, we'd like to fit the MLR model in *R*, which needs the package "nnet" to be installed.

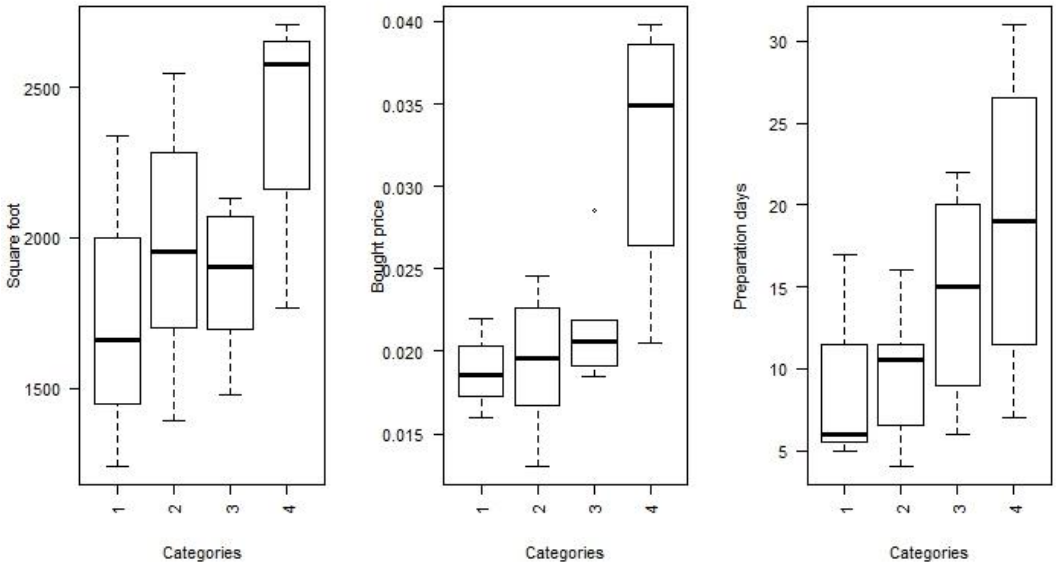

Figure 18: Box plots of predictors w.r.t. categories

In the model, we have 20 independent observations in total with 4 explanatory variables, including square foot, bought price, preparation days and quarter. To construct the logits in the multinomial case, one of the m ($m = 4$) categories is considered the base level and all the logits are constructed relative to it.[7] Any category has the possibility to be regarded as the base level and be labeled m since the categorical variable is unordered. Let $\pi_1, \pi_2, \ldots, \pi_J$ denote the multinomial probability of an observation falling in the jth category [7], with $\pi_1 + \pi_2 + \cdots + \pi_J = 1$. In order to discover the relationship between the respective probabilities and the predictor variables, the general formula of Multinomial Logistic Regression reports the log odds as follows:

$$log\left(\frac{\pi_j(x_i)}{\pi_m(x_i)}\right) = \beta_{oj} + \beta_{1j}x_{1i} + \cdots + \beta_{mj}x_{mi},$$

where $j = 1, 2, 3$ and $i = 1, 2, \ldots, 20$.

In our situation, we want to the 4 categories of listed to sold days based on 4 explanatory variables. The "LSD < 30" (level 1) has been chosen as the base level. As a result, the MLR would take the form below, where $i = 2, 3, 4$, referring to the above equation,

$$log\left(\frac{P(LSD=i)}{P(LSD=1)}\right) = \beta_o + \beta_1(sqft) + \beta_2(bought\ price) + \beta_3(prep.days) + \beta_3(quarter).$$

After modeling in R, we removed the predictor "quarter" since it always showed insignificance in the models due to lack of information. Therefore, we eventually obtained the output of MLR in Figure 19.

Model : multinom(Dummy ~ Prep.Days + square.foot + Bought.For, censored=FALSE) **Residual Deviance**: 27.28789 **AIC**: 51.28789				
Coefficients	(Intercept)	Prep.Days	square.foot	Bought.For
2	-2.253	-0.003	0.002	-0.002
3	-5.979	0.190	0.000	0.022
4	-223.3	2.054	0.039	0.400
Std. Errors	(Intercept)	Prep.Days	square.foot	Bought.For
2	0.006	0.150	0.002	0.018
3	0.017	0.162	0.002	0.019
4	0.016	1.785	0.017	0.100
Z-test	(Intercept)	Prep.Days	square.foot	Bought.For
2	-356.9	-0.018	0.943	-0.114
3	-355.2	1.172	0.008	1.159
4	-14077.3	1.151	2.237	3.957
p-value	(Intercept)	Prep.Days	square.foot	Bought.For
2	2	1.014	0.346	1.091
3	2	0.241	0.994	0.247
4	2	0.250	0.025	0.000
Odds	(Intercept)	Prep.Days	square.foot	Bought.For
2	0.105	0.997	1.002	0.998
3	0.0025	1.210	1.000	1.022
4	0.000	7.800	1.040	1.491

Figure 19: Multinomial logistic regression output

From the results, we want to interpret in the following features.

- **Coefficients.** With regard to the formula of MLR, these are the logit coefficients relative to the reference category. For instance, considering the preparation days, the -0.003 indicates that for one day increase during the preparation term, the logit coefficient for level 2 (when listed to sold days are from 1 to 2 months) relative to level 1 (listed to sold days are within a month) will go down by 0.003. In other words, if the preparation days increase by one unit, the probability of house being sold within a month will be higher compared to level 2. However, this conclusion is opposite to level 3 and 4. And

the variable of bought prices suggests similar conclusion to the preparation days. However, the coefficients under "square foot" are all greater than zero (the value 0.000 is also larger than zero), which means that if the floor size increases by one unit, the probabilities of staying in level 2, 3, and 4 are relatively higher than staying in the base level 1.

- **Z-test and p-value.** The test statistic *z* is the ratio of the coefficients to the standard errors. The p-value suggests strong significance at the level of 5% when the value of a variable is less than 0.05 so that the null hypothesis can be rejected and the estimated parameter could be significant [16]. Unfortunately, the values shown in Figure 19 cannot be concluded to have great significance in general. For example, for level 2 relative to level 1, the *z* test statistic for preparation days is equal to -0.018 with an associated p-value of 1.014. Then, setting the significance level to be 0.05, we fail to reject the null hypothesis so that the regression coefficient of preparation days cannot be statistically different from zero given the square foot and bought prices are also in the model. Neverthless, we cannot say the explanatory variables are insignificant, since the multinomial logistic regression model might perform better if we have more observations.
- **Relative risk.** The ratio of the probability of choosing one outcome category over the probability of choosing the baseline category is often referred to as relative risk, which is sometimes referred to as *odds* [16] as well. The relative risk is the right-hand side linear equation exponentiated, leading to the fact that the exponentiated regression coefficients are relative risk ratios for a unit change in the predictor variable.[16] We can exponentiate the coefficients to see the risk ratios. For example, the relative risk ratio for one unit increase in bought price is 0.998 of level 2 compared to the base level 1.

Overall, the multinomial logistic regression is definitely a good approach to analyze the effects on a multiple (greater than 2) unordered categorical response variables. Therefore, we can still do further modeling if more data set is provided in the future.

Chapter 4

INTRO TO SURVIVAL ANALYSIS

Survival analysis is generally defined as a set of methods for data analysis where the outcome variable is the time until the occurrence of an event of interest, where the event can be death, occurrence of a disease, marriage, divorce, and etc. The time to event or survival time can be measured in days, weeks, years, and so forth.[18] However, the reason why linear regression is not a good choice to be applied to model the survival time as a function of a combination of explanatory variables is that, firstly, because of the restriction that survival times are typically non-negative numbers and have skewed distributions with long tails, they should be transformed in order to perform the ordinary linear regression; secondly, the linear regression cannot effectively handle the censoring of observations.[18] The set of data is called censored (including right censored and left censored) when the information of their survival time is not known completely, which is important to indicate the fact of missing data. The data being right censored occurs when the event of interest didn't occur during the study or occurred after the end of the study so that the continued information cannot be recorded. Inversely, the data being left censored happens when the survival time commenced before the study began so that the existing information cannot be recorded either.

In the study of survival analysis with respect to the project, we want to consider the bought to sold days (the period from bought dates to sold dates) to be the survival time. However, the original data set does not have complete information of sold dates as is mentioned in the introduction of the data set. We treated the completed data (by filling up the information from Zillow) as censored for the survival analysis. And the dependent variable consists of two aspects: one is the time to event

which is the bought to sold days, and the other is the event status, which records if the event of interest occurred in the original data set (status = “1”) or not (status = “0”). In the analysis, we are able to estimate two functions depending on time, which are survival and hazard functions. Survival function suggests the probability of surviving or not experiencing the event in the duration of study up to that time; while the hazard function suggests the potential that the event will occur, per unit of time, given that an individual has survived up to the specified time.[18] So, the purpose of survival analysis in this section is to describe the relationship of a factor of interest to the time in days to the “bought.sold days”, in the presence of three covariates, such as square foot, bought prices and quarter. In this way, there are some models to generate the relationship of a set of predictors with the survival time. Methods include parametric, nonparametric, and semiparametric approaches.[18]

Parametric methods assume that the underlying distribution of the survival times follows certain known probability distributions, such as Gamma, exponential, Weibull and etc. With regard to the bought to sold days, the histogram with a red curve of distribution is shown as Figure 20. We might be able to determine the probability distribution as the Gamma. However, Gamma parametric has broken for all three covariates, which would be discussed in the following section.

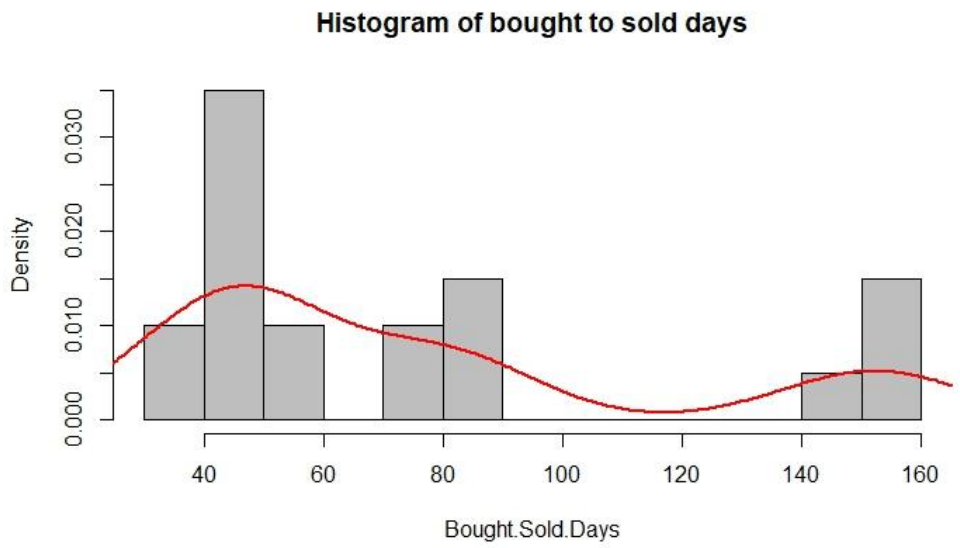

Figure 20: Histogram of bought to sold days

Nonparametric method, also known as the Kaplan Meier method, is used to estimate the survival probability from observed survival times [19] (Kaplan and Meier, 1958). It mainly tests for overall differences between estimated survival curves of two or more groups of subjects, such as females versus males, or quarter 3 versus quarter 4 in our project. However, it is not suitable to be used if adding other numerical variables such as square foot and bought prices.

A more popular regression model for the survival analysis is called Cox proportional hazards model [20] (Cox, 1972), which is a semiparametric model [13] in which dependence on the explanatory variables is modelled explicitly but no specific probability distribution is assumed for the survival times. It will be discussed in details in the following section.

4.1 Cox proportional-hazards model

The Cox proportional hazards regression model is commonly used in medical research for investigating the association between the survival time of patients and one or more predictor variables.[9] It works for both quantitative predictor variables and for categorical variables.[9] The purpose of the model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g. infection, death) at a particular point in time. This rate is referred to as the "hazard rate", and the predictor variables (or factors) are usually termed covariates in the survival analysis literature.[9]

The Cox regression is expressed by the hazard function which is denoted by $h(t)$ in the following formula. Briefly, $h(t)$ can be interpreted as the risk of dying at time t. The formula is as follows:

$$h(t) = h_0(t) \times exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)$$

where t is the survival time, $h(t)$ is the hazard function determined by the set of covariates $(x_1, x_2, \ldots, x_n)$, the coefficients β_i measures the impact (i.e. the effect size) of covariates, and the $h_0(t)$ is the baseline hazard corresponding to the value of hazard when all the x_i's are equal to zero (i.e. *exp* term equals 1). The 't' in $h(t)$ indicates that the hazard may vary from time to time. The quantities $exp(\beta_i)$ are called hazard ratios (HR); a value of β_i larger than zero or equivalently a hazard ratio larger than 1 indicates that the hazard of the event will increase and the length of survival will decrease as the covariate x_i increases.[21] In other words, a hazard ratio above 1 suggests a covariate is positively associated with the event probability, and therefore negatively associated with the length of survival. However, a hazard ratio equal to 1 indicates no effect.

To begin the modeling and analysis in *R*, we need to install two packages: *survival* (for computing survival analysis), and *survminer* (for visualizing survival analysis results). Considering the "Bought.Sold.days" as the outcome, we used the floor sizes of houses, bought prices, and calendar quarters as the predictor variables. The syntax is *coxph(Surv(time = Bought.Sold.days, status) ~ square.foot + Bought.For + Quarter)*. The summary of Cox proportional-hazards model is shown in Figure 21, which is interpreted below.

```
                      coef  exp(coef)   se(coef)      z Pr(>|z|)
square.footage -5.084e-05  9.999e-01  9.123e-04 -0.056   0.9556
Bought.For     -1.889e-02  9.813e-01  1.098e-02 -1.720   0.0854 .
QuarterQ4      -9.479e-01  3.876e-01  6.808e-01 -1.392   0.1639
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

               exp(coef) exp(-coef) lower .95 upper .95
square.footage    0.9999      1.000    0.9982     1.002
Bought.For        0.9813      1.019    0.9604     1.003
QuarterQ4         0.3876      2.580    0.1021     1.472

Concordance= 0.699  (se = 0.101 )
Rsquare= 0.402   (max possible= 0.932 )
Likelihood ratio test= 10.3  on 3 df,   p=0.02
Wald test            = 4.7  on 3 df,   p=0.2
Score (logrank) test = 7.54  on 3 df,   p=0.06
```

Figure 21: Output of Cox proportional-hazards model

- **Statistical significance.** The column marked "*z*" (*z* = *coef*/*se*(*coef*)) in the summary provides the Wald statistic value. The Wald statistic [9] evaluates, whether the β_i coefficient of a given variable is statistically significantly different from 0. From the results, the p-value of "bought prices" is relatively small compared to the other two predictors, which means this variable has a statistically significant coefficient. However, the "square footage" and "quarter" are not statistically significant enough, since there p-values are relatively large.

- **Regression coefficients.** The second important feature in the Cox model results is the sign of the regression coefficients (which is "coef" in the summary). Specifically, a positive sign means that the hazard is higher, and thus the prognosis worse [21]. In this model, all of the predictor variables have negative signs as coefficients, which indicates a lower risk to the bought-sold days. As for the quarters, Winter provides a smaller risk to the bought-sold days than Autumn.
- **Hazard ratios.** The exponentiated coefficients ("*exp*(coef)"), also known as hazard ratios (HR), are defined as the measure of an effect of an intervention on an outcome of interest over time [21] (or give the effect size of covariates). As for the measurement values, we noticed that the hazard ratios for square foot and bought prices are close to and slightly smaller than 1, which implies the hazard will decrease, or we can say that the square foot does not have any important effect on the bought to sold days in this model (if regarding 0.9999 just as 1). In other words, the "square foot" and "bought prices" make little contribution to the bought-sold days of house sale, although it might not be true in reality. Since the hazard ratio of Quarter is much smaller than 1, it indicates a strong relationship between houses sale in Winter and will decrease the risk of bough-sold days.
- **Confidence intervals of the hazard ratios.** The summary output also gives upper and lower 95% confidence intervals for the hazard ratio, which shows consistent conclusions to the hazard ratios; except that Winter could also increase the risk sometimes.
- **Global statistical significance of the model.** Finally, the p-values can never be ignored in the model, which are for three alternative tests for overall significance of the model: the Likelihood-ratio test, Wald test, and Score (logrank) statistics. These three methods are asymptotically equivalent and

give the same results for large sample size. However, for small sample size *N* (*N=20* in our set of data), they may differ somewhat, while the Likelihood ratio test has better behavior, which should be generally preferred. Overall, the p-value of Likelihood ratio tests is less than 0.05, which affirms the statistical significance of the model in a way.

Next, we did a *cox.zph*() test to check for the proportional-hazards (PH) assumption. From the output below (Table 4), the p-values for each predictor variables as well as the global test are larger than 0.05, which suggests that the test is not statistically significant for each of the covariates. And the global test is not statistically significant either. Therefore, the proportional hazards could be assumed. In addition, since the *cox.zph*() test utilizes the Schoenfeld residuals against the transformed time, the large p-values implies that there are no time dependent coefficients to be cared about.

Table 4: *cox.zph*() test of Cox model

	rho	**chisq**	**p**
square.foot	-0.0671	0.0227	0.880
Bought.For	-0.6262	2.0429	0.153
QuarterQ4	-0.4969	2.2376	0.135
GLOBAL	NA	4.0986	0.251

Moreover, we also performed the graphical diagnostic plots by using the function *ggcoxzph*() (included in the "*survminer*" package), which applied the figures of the scaled Schoenfeld residuals against the transformed time for each covariate. In Figure 22, the solid line is a smoothing spline fit to the plot, while the dashed lines stand for a +/- 2-standard-error band around the fit [21]. Obviously, we can see that

there is no pattern with regard to time, which is consistent to what we analyzed for the p-values.

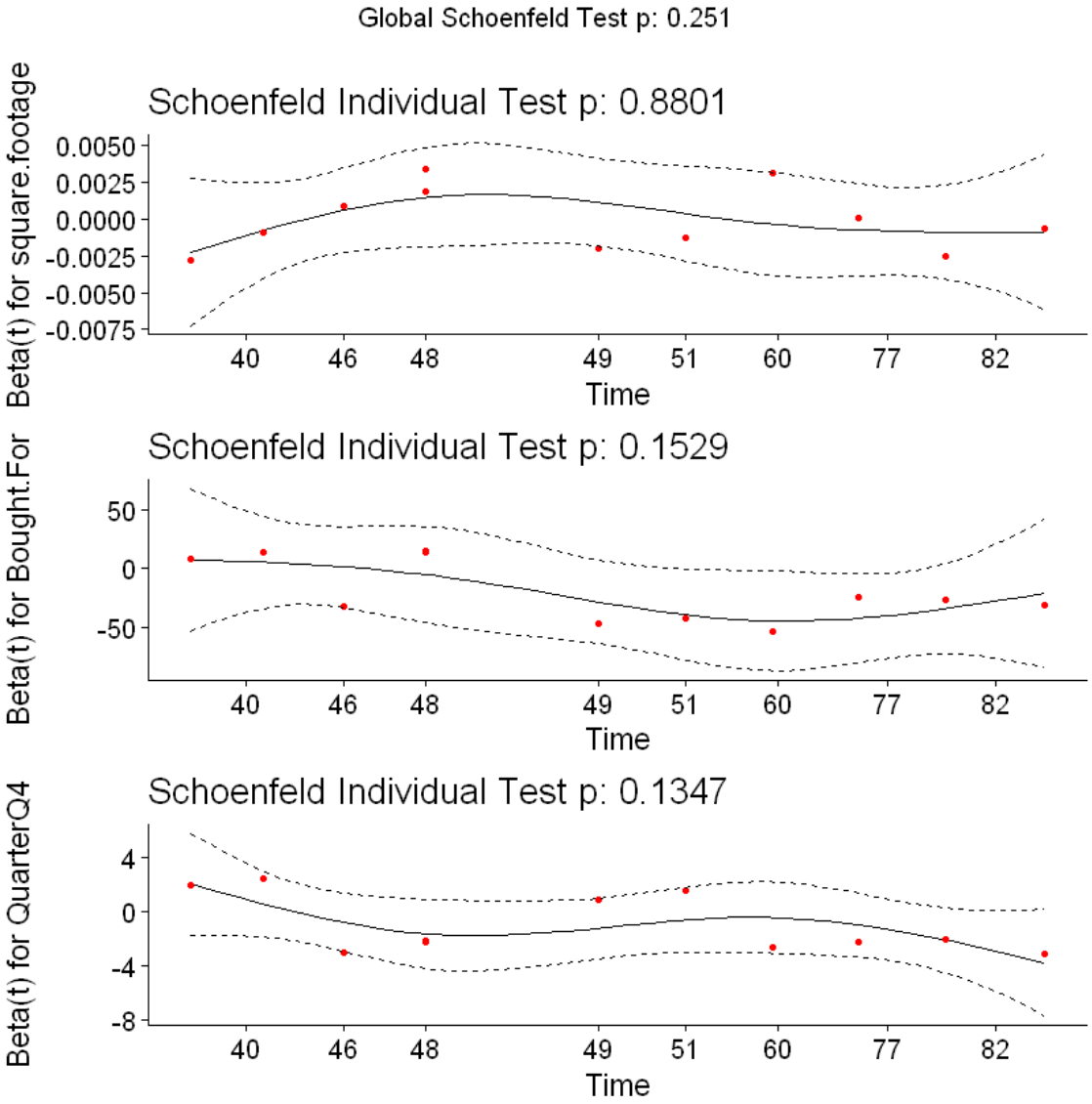

Figure 22: Diagnostic plots – Cox PH model

Next, we could also test influential observations or outliers by checking either "dfbeta" values or "deviance residuals", which can be accomplished by the function *ggcoxdiagnostics*() from *survminer* package. "dfbeta" gives the estimated changes in

the regression coefficients upon deleting each observation in turn, as well as the coefficients divided by standard errors. The result is shown in Figure 23.

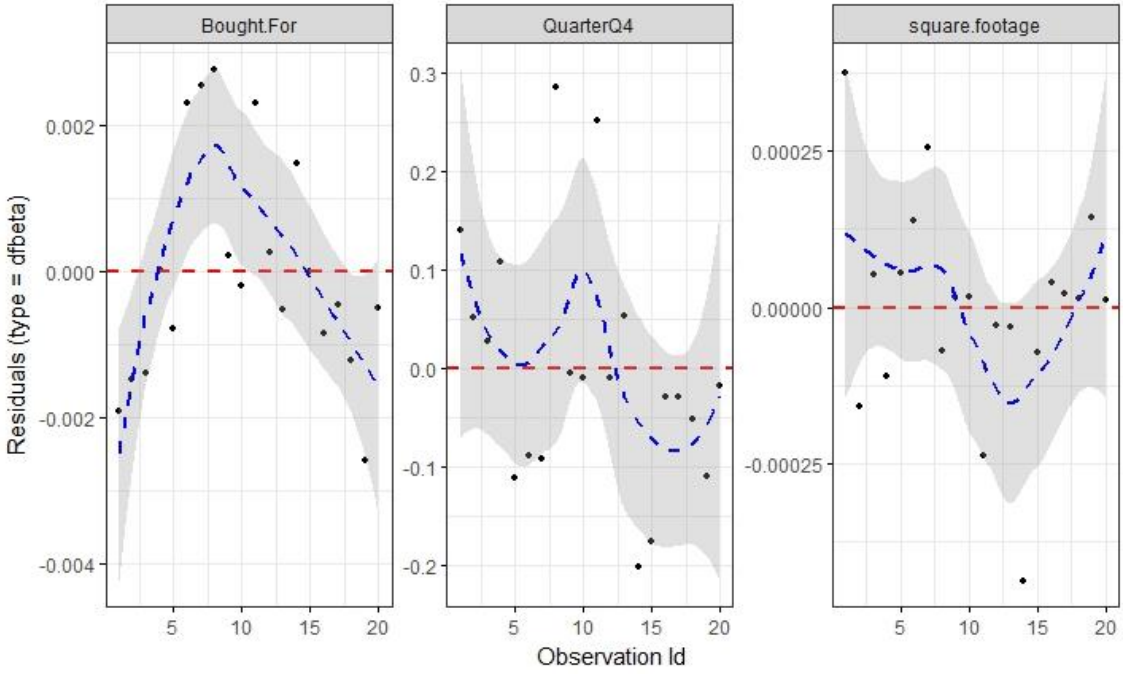

Figure 23: Diagnostic plots – dfbeta

Unfortunately, since the diagnostic plots indicate some problematic patterns, we might guess that the observations are terribly influential individually. We want to check the outliers by using the deviance residuals which are the normalized transform of the martingale residual. Ideally, these residuals should be roughly symmetrically distributed around zero with a standard deviation equal to 1. However, the result shown in Figure 24 is not ideal, since the patterns make it apparent that the larger or small outliers are poorly predicted by the model.

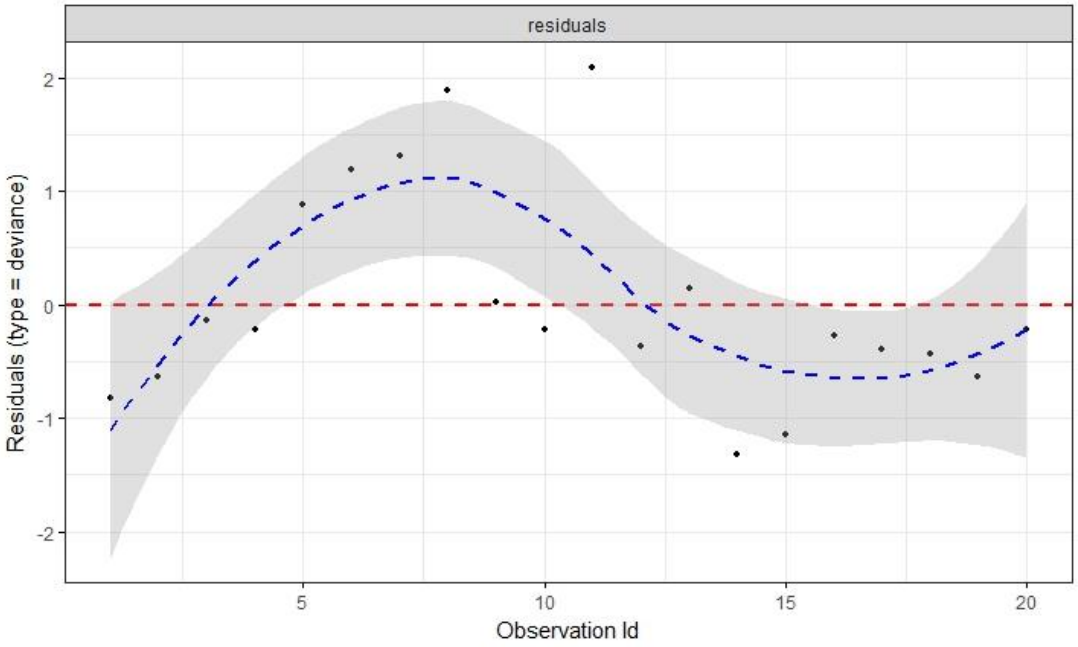

Figure 24: Diagnostic plots – deviance

4.2 Gamma parametric survival model

The Cox model for survival data is ubiquitous in medical research, since the effects of predictors can be estimated without needing to supply a baseline survival distribution that might be inaccurate. However, fully-parametric models have many advantages, and even the originator of the Cox model has expressed a preference for parametric modelling [23] (see Reid 1994). Fully-specified models can be more convenient for representing complex data structures and processes [24] (Aalen et al. 2008), e.g. hazards that vary predictably, interval censoring, frailties, multiple responses, datasets or time scales, and can help with out-of-sample prediction.[22]

As is mentioned, we would choose the Gamma parametric survival model since the survival time presented a rough Gamma distribution. '*flexsurv*' is an *R* package

for fully-parametric modeling of survival data. To maintain the same covariates as the Cox model, the syntax in this case became *flexsurvreg(Surv(time = Bought.Sold.days, status) ~ square.foot + Bought.For + Quarter)*. However, the model broke with these 3 covariates, since optimizing with *BFGS* requires the gradient of the function being minimized.

Chapter 5

LIMITATION AND FUTURE WORK

The most important limitation that concerns us is the lack of evidence and observations from the given set of data since the data set is quite small. It is restrictive to the analysis of discussed models since sometimes we could not negate an appropriate model which didn't produce statistically significant results, for example, the logistic regression models might fit better if more information from data can be applied; reversely, we couldn't negate a false model due to relatively good fitting, for example, the simple linear model may not be satisfied all the time in the real world. On the other hand, there are some other terms of variables that didn't show good adequacy, such as the interaction variables constructed from "Quarter × Preparation days" and "Quarter × Square footage". There are also some models that cannot be generated for the given observation, like random effects model with regard to the zip code. In addition, since the business of house sale does not always remain unchanged with regard to quarters, we can probably discover some patterns in quarters if we have more information to discuss. For instance, summer might be a more popular season to sell houses than winter.

To be detailed, there are three sections we did in the project but didn't get ideal outcomes, which are worthwhile to be discussed as follows.

5.1 Mapping

As is mentioned in the introduction, iBuyers don't actually buy all homes, so is there any feature of communities that are preferred by iBuyers?

Before modeling, we plotted the Georgia map using the *maps* and *ggplot2* packages installed from *R* and added points of houses with their longitudes and latitudes, where the points have different symbol sizes that are proportional to profit per square foot as shown in Figure 25, in order to see some patterns or features around the communities along with an enlarged map in the top right corner. Unfortunately, referring to the satellite map from Google, we didn't find any feature in the specific community and the each house is quite far away from others in reality. In this case, we cannot say there is any preference for Opendoor to choose houses regarding to good communities.

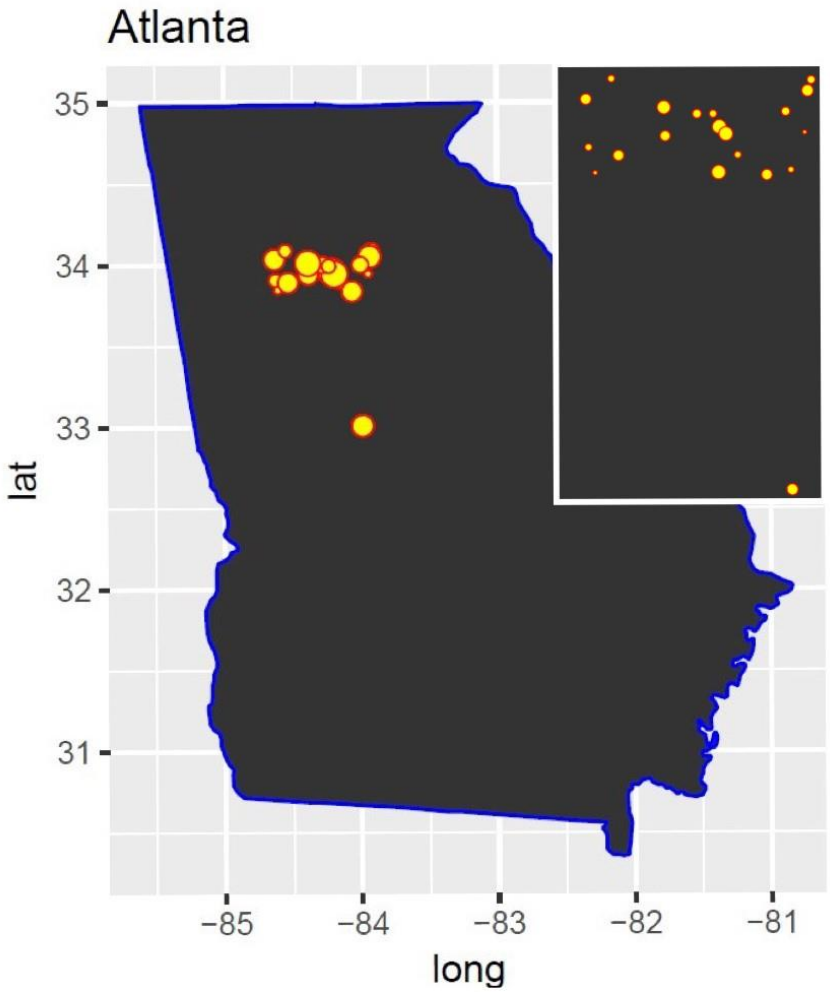

Figure 25: Map of observations in Atlanta

Nevertheless, if it is possible to obtain more data resources in the future, there might be some features that we can extract just by looking at the maps, such as whether or not the houses are close to busy roads/bus stations/parks, whether or not the elementary/middle/high schools are of high quality, and so forth. These will provide much information for statistical modeling in the long run.

5.2 LMMs - Random effects model

Linear Mixed Models (LMMs) are an extension of simple linear regression models to allow and incorporate both fixed and random effects, which are particularly used when there is non independence in the data, such as arises from a hierarchical structure.[17] For example, students could be sampled from within different majors. A fixed effect is a parameter that does not vary, while a random effect is a parameter that is a random variable. Observations can be correlated in the random effects model.

In our project, in order to describe and predict the effects on the profit per square foot by the linear mixed effects model, we would consider the zip code as a random variable, since there are several different zip codes for each house. To perform the mixed effects model in *R*, we need to use the *lmer* function from the *lme4* package. The random variable can be used in the form "(1 | zip.code)" which indicates that there is a random effect for each zip code and this effect is nested within the intercept (the whole model). Then, we can choose square footage and preparation days to be the fixed effects. As a result, the output of the mixed effects model is shown as below.

Model: lmer(prof.sqft ~ square.foot + Prep.Days + (1 | zip.code))			
Random effects:			
Groups	Name	Variance	Std. Dev.
zip.code	(Intercept)	0.000	0.00
Residual		6.971	2.64
Fixed effects:	Estimate	Std. Error	t value
(Intercept)	21.549	2.9109	7.403
square.foot	-0.0077	0.0016	-4.843
Prep.Days	0.2883	0.0948	3.040

Figure 26: Output of Random Effect Model

As for the coefficients of random effects, the column of standard deviation stands for the variability of the random effect added into the model. However, the random item "zip code" has a zero variability, as does the value of variance. This seems to be of insignificance, since the zip code differs from each house for 20 observations in total. In this way, the random effects model is not meaningful for the particular data set. In addition, the "residual" represents the variability that is not because of zip code, which corresponds to the error term. The coefficients of fixed effects imply their influence on the profit per square foot. For example, the coefficient of preparation days is 0.2883, which means that, to get more profit per square foot, Opendoor has to increase the preparation days by 0.2883.

To confirm that the random term is not significant in this model, we used an ANOVA-like test called *rand* function from the *lmerTest* package to measure the random effect, "zip code" in the model. The p-value is found to be 1. Therefore, the zip code is not needed in the model.

However, the linear mixed effect model is an appropriate approach to choose so that the random variables can also be analyzed and compared to the simple linear model. Therefore, if we have more data set in the future, and some of the zip codes

are repeated, we can try the random effects model again to discuss the more significant model.

Chapter 6

CONCLUSION

The project is structured around studying the data set of house sales by Opendoor in Atlanta. Opendoor, which is known as an iBuyer, buys houses from customers and re-lists them for sale with the goal of doing so at a quick rate compared to traditional house agency. This article presents different kinds of models in mathematical statistics, in order to describe and predict the effects not only on the profit per square foot, but also on the listed to sold days, as well as the bought to sold days. The models were performed in the following three important sections.

Firstly, we started with the simple linear regression model to generate the effects on the quantitative response variable, “profit per square foot”, with only two predictors, “square foot” and “preparation day”, since the variable of quarter didn't show statistical significance to the response. This model fit well although the diagnostic plots indicated slightly problematic patterns due to the residual versus fitted value plot. We used the log-transformation as well as the square root transformation on the preparation days manually, however, the results were not obviously improved. To extend the simple linear model, we tried the Generalized Linear Model (GLM) with a Gamma regression, with which we obtained better diagnostic plots, which indicates the predictors are more significant to fit the GLM. Although the AIC and BIC values had not approved, the difference between the two models can be ignored.

Considering the non-linearity appeared in the linear regression models, we performed the Generalized Additive Model (GAM) which allows for non-linear relationships between each predictor variable and the response by using smoothing

functions. In this case, we got a best GAM model that used a linear function of square foot and a smoothing function of the preparation days.

Secondly, we selected the variable of listed to sold days and converted it into binary or binomial variables, which was treated as the qualitative/categorical response; the predictors used were square foot and bought prices. The listed to sold days were firstly binary, since, for example, the binary response is equal to 1 if the days are less than 30, and equal to 0 if the days are larger than 30. Because the days range from less than 1 month to larger than 4 months, we performed 4 logistic regression models. However, the results were not good enough due to lack of observations. Furthermore, since GAM can also be used for categorical response variables, we found a non-linear relationship between the bought prices and the listed to sold days. Indeed, the GAM logistic regression worked better than the general logistic regression.

On the other hand, to consider the categorical response as multiple (more than 2), we created 4 groups of the listed to sold days separated by months into 4 levels: level 1 corresponds to when the days are within a month, level 2 when the days are from 1 month to 2 months, and so forth; the multinomial logistic regression was generated in this case, which is an extension of the logistic regression model. Overall, the multinomial logistic regression is definitely the best approach in this situation since the categories of response are larger than 2, although we still need to do further modeling if more data is provided in the future.

Finally, we used the survival analysis to discuss the risk of square foot, bought prices, and quarters to the bought to sold days since the survival analysis is generally defined as a set of methods for data analysis where the outcome variable is the time until the occurrence of an event of interest. The survival time is the bought to sold days. Survival analysis includes three kinds of methods, parametric, nonparametric, and semiparametric. Only the semiparametric method is appropriate in our problem,

which is represented by the Cox proportional hazard regression model. Unfortunately, as a result, the Cox model didn't produce good results since we only have 11 observations for survival analysis, the observations are terribly influential individually, and since some outliers are poorly predicted by the model. Even though the Cox model does not seem to be suitable in our situation, it's still worthwhile to use the model in further study.

BIBLIOGRAPHY

[1] Dalrymple, Jim. "What is an iBuyer?" *Inman*, Postamo Social Media Corporation, 10 December 2018, www.inman.com/2018/12/10/the-essential-guide-to-ibuyers/.

[2] Schafer, Ricardo. "The Rise of the Instant Home Buyer Model." *Medium*, A Medium Corporation, 29 August 2018, medium.com/loric-ventures/the-rise-of-the-instant-home-buyer-model-3c527811acc7.

[3] Read, Cortney. "What is an iBuyer?" *Offerpad*, Offerpad Corporation, 21 December 2017, blog.offerpad.com/what-is-an-ibuyer/.

[4] "WHAT'S THE DIFFERENCE BETWEEN IBUYING AND HOME FLIPPING?" *Inman*, Postamo Social Media Corporation , 10 December 2018, www.inman.com/2018/12/10/the-essential-guide-to-ibuyers/.

[5] "WHAT KIND OF CUSTOMERS ARE USING IBUYERS?" *Inman*, Postamo Social Media Corporation, 10 December 2018, www.inman.com/2018/12/10/the-essential-guide-to-ibuyers/.

[6] Delprete, Mike. "Inside Opendoor: what two years of transactions say about their prospects?" *MD*, 13 December 2016, www.mikedp.com/articles/2016/12/13/inside-opendoor-what-two-years-of-transactions-say-about-their-prospects.

[7] Chatterjee, Samprit, and Ali S. Hadi. *Regression analysis by example*. Wiley, 2015.

[8] Cox, David Roxbee. *Regression models and life-Tables*. Journal of the Royal Statistical Society. Series B (Methedological), Vol. 34: 187–220, 1972.

[9] Easy Guides, "Cox Proportional-Hazards Model." *R-bloggers*, R-bloggers Corporation, 12 December 2016, www.r-bloggers.com/cox-proportional-hazards-model/.

[10] "What is Null and Residual deviance in logistic regression." *Analytics Vidhya*, Analytics Vidhya Corporation, August 2015, discuss.analyticsvidhya.com/t/what-is-null-and-residual-deviance-in-logistic-regression/2605.

[11] "Generalized linear Model." *Wikipedia: The Free Encyclopedia.* Wikipedia, The Free Encyclopedia, 5 April 2019, en.wikipedia.org/wiki/Generalized-linear-model.

[12] Nelder, J.A. and Wedderburn, R.W.M. *Generalized linear models*. Journal of the Royal Statistical Society, Series A, 135, 370-384, 1972.

[13] Dobson, Annette J. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2002.

[14] Hastie, T.J. and Tibshirani, R.J. *Generalized Additive Models*, Chapman and Hall, London, 1990.

[15] James, G., Witten D., Hastie T.J. and Tibshirani, R.J. *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistics, DOI 10.1007/978-1-4614-7138-7-4, Springer Science+Business Media New York, 2013.

[16] "Multinomial Logistic Regression | R data analysis examples." *IDRE Stats,* stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/.

[17] "Introduction to Linear Mixed Models." *IDRE Stats,* stats.idre.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/.

[18] Despa, Simona. "What is Survival Analysis?" *Cornell University Cornell Statistical Consulting Unit,* www.cscu.cornell.edu/news/statnews/stnews78.pdf.

[19] Kaplan, E.L., and Meier, Paul. *Nonparametric estimation from incomplete observations*. Journal of the American Statistical Association 53: 457–481, 1958.

[20] Cox, David Roxbee. *Regression models and life-Tables*. Journal of the Royal Statistical Society. Series B (Methedological), Vol. 34: 187–220, 1972.

[21] "Cox Proportional-Hazards Model." *STHDA,* www.sthda.com/english/wiki/cox-proportional-hazards-model.

[22] Jackson, Christopher. "flexsurv: A Platform for Parametric Survival Modelling in R." *MRC Biostatistics Unit, Cambridge, UK,* cran.r-project.org/web/packages/flexsurv/vignettes/flexsurv.pdf.

[23] Reid, Nancy. *A Conversation with Sir David Cox*. Statistical Science, **9**(3), 439-455, 1994.

[24] Aalen, Odd O., et al. *Survival and Event History Analysis: A Process Point of View*. Springer-Verlag, 2008.