Research Questions (Steven Spiegel)

- **Can we make a linear regression model to accurately predict Housing Sale Price?** (chosen)
- Is there any significant difference in quality of housing and Lot Size, roof type, or type of exterior?
- Does distance to schools affect housing price?

Ames Iowa: Prediction House Prices using Linear Regression of internal and external variables.

Abstract

This paper seeks to predict house prices using Multiple Linear Regression using various internal and external variables found in De Cock (2011). The Ames Housing Dataset contains 2930 housing records with 80 explanatory variables from the Ames, Iowa Assessor's Office. The question we want to ask is: Can we reasonably predict house prices only using a few fields that most people could easily access. These variables include internal ones such as ground floor area, basement floor area, overall condtion, etc, and external ones such as distance to a school and neighborhood. We will explore what variables play a bigger factor in house prices and which ones play a smaller factor. Then we will see how well our model predicts the house sales price.

Introduction

Home prices are an important factor in a society's overall economic health. Historically, purchasing a home gives an individual tremendous economic advantage over a rentor, since the home buyer can build equity into their property. The housing market is a societal and economic indicator by showing how that particular society values homes and home ownership. Indeed, according to Olga and Antonios (2019), the housing market can provide key insights into the economic health of that society beyond normal banking indicators. Hence, the ability to accurately predict the price of a house will allow the prospective buyer to better understand what attributes of a house are most valuable, and also what attributes of a home contributes most to its cost.

Modelling home prices is not a trivial task. One must have an understanding of various factors that affect the housing market and what aspects of a home can affect its price. These can include environmental factors, locations to schools, crime rate, etc. Accurately modeling these factors to compute home sale prices often requires the use of more advanced statistical and machine learning techniques. Truong et al. (2020) implemented several machine learning algorithms to predict home prices. These included a Random Forest classifier, extreme gradient boosting, light gradient boosting, and a hybrid between several models. The model that achieved the greatest accuracy was a hybrid regression approach, which implemented a combination of machine learning frameworks. This dataset had over 200,000 records and was split into a training and test set to get an accurate measure of model performance.

Mora-Garcia, Cespedes-Lopez, and Perez-Sanchez (2022) performed similar models in their study using random forest regressor, gradient boosting, extreme gradient boosting, and linear regression. The paper demonstrated that all machine learning algorithms were superior in their prediction to that of linear regression. They also required a large amount of fine tuning, feature engineering, and preprocessing of the data.

Zhang (2021) used a simpler method with multiple linear regression models. First, they computed the *Spearman correlation coefficient* (denote ρ_s) in order to ascertain which features in their dataset correlated strongly with housing prices. The values of $\rho_s \in [-1, 1]$, where -1 and 1 are perfect negative and positive correlations, respectively. For any set of observations X and Y where $x \in X$ and $y \in Y$ can be ranked (ordered), The equation for the pearson coefficient is

$$\rho_s = \frac{6 * \sum d_i}{n(n^2 - 1)} \tag{1}$$

where d_i is the difference in rank of values X_i and Y_i, and n is the number of observations. Using this equation, they were able to find the attributes that had higher correlations (i.e. rankings closer to -1 or 1) and performed linear regression model on these attributes.

While it can be demonstrated that using more advanced machine learning algorithms is superior to that of linear regression, these types of algorithms require specialized knowledge of machine learning and hyperparameter tuning that are beyond the scope of this study. In fact, hyperparameter tuning is in and of itself an entire area of study. Furthermore, as seen in Truong et al. (2020), the amount of data available was on the order of 10^5, allowing for the use of more advanced machine learning algorithms and splitting between training and testing. These papers also included data outside the data provided in the Ames dataset and may be something that is not easily reached by a regular person trying to get a good idea of housing prices and sales. Hence, we will only be using a few of the datapoints provided in the Ames dataset, simulating someone who has limited access to data might be able to perform.

The question we want to answer is, how well does a linear regression model work in predicting housing prices using only a few variables? The benefit of this is most people will be able to easily perform this task without knowing anything about hyperparameters, feature engineering, etc. In fact, it could be done using a prebuilt Excel spreadsheet that most could access. Linear regression is relatively easy to understand and can be done very quickly without too much parameter tuning.

Methods

Internal house factors

We will first see what internal factors contribute most to the overall sales price of the house. These factors are intrinsic to the home itself rather than external such as school district, disntance to school, neighborhood, etc. It should also be noted that these fields are easily retrieved from most assessor data and will also be important to home buyers. While locations of homes may vary, many of these internal factors still play an important role in the sale of a house.

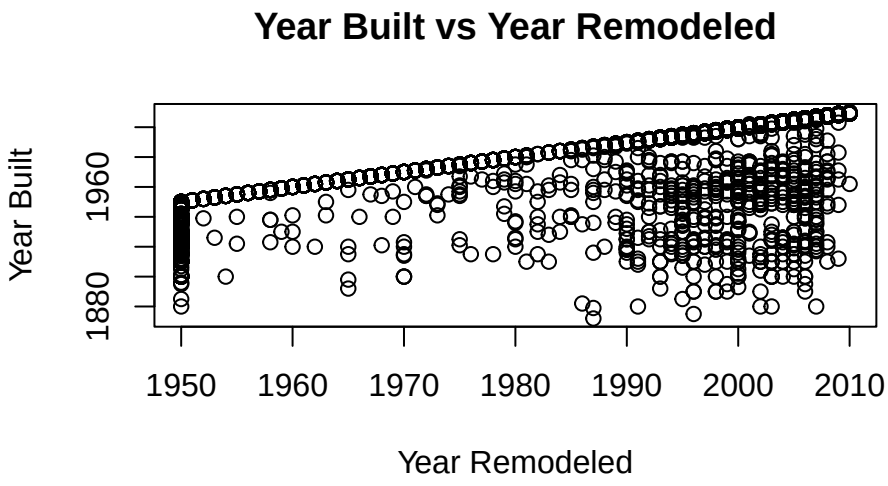

Figure 1: Year Renovated vs year Built

Outliers

First we will look at any outliers within the fields. Normally, we would not remove any outlier data if we don't have a good reason to. However, from De Cock (2011), it is recommended to remove homes that have

more than 4000 square feet of area (see De Cock (2011)), as they are true outliers and don't reflect market values at the time. We will keep all other data values, as there is no apparent reason for them to be removed.

Additionally, we will also only consider homes that were sold after 1950. If we examine Figure 1, we see that houses built before 1950 were never renovated until 1950. This renovation was likely due to mandatory updates from the city. This will skew the data since this is not a renovation from the owner but rahter a mandatory one.

Simple Linear Regression

We will use a simple linear regression model to find the best fit linear equation that satisfies the following:

$$y_i = W_0 + W_1 x_i + W_2 x_i + \dots W_n x_i + e_i \tag{2}$$

Where x_i, y_i are the observations, $W_0 \dots W_n$ are the weights, and e_i is the error residual between the predicted value of $W_0 + W_1 x_i + W_2 x_i + \dots W_n x_i$ and y_i. If we include a bias value for x_i (a value of one for each x), we can rewrite the equation as:

$$y_i = W_0 1 + W_1 x_i + W_2 x_i + \dots W_n x_i + e_i \tag{3}$$

Where W is the $1\text{X}(n+1)$ weight matrix. In this case, our predicted value, y, is the sale price of the home and our x are the various internal factors we wish to investigate. While there are many various internal factors we could consider, the ones we picked are ground living area, roof style, total basement area, total garage area, Age, and Remodel Age.

Determining a model's prediction ability

Determining how well a model predicts housing prices is not a trivial task. We may be able to find the optimal weights analytically, but that does not mean we will get a good result, especially if the data is non-linear. However, for linear regression, we can use the adjusted R^2 statistic, which is the level of variance in the observations that is accounted for by our model. Specifically, we use the adjusted R^2, which is used for multiple linear regression since it takes the number of degrees of freedom into account as well. Equation 4 and 5 show how R^2 and adjusted R^2 (labeled $\tilde{R}^2$) are calculated, respectively.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{4}$$

$$\tilde{R}^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1} \tag{5}$$

where N is the number of samples and p is the number of explanitory variables.

We will also look at the residuals of the model. A residual for observation y_i and prediction $\hat{y}_i$ is

$$e_i = y_i - \hat{y}_i \tag{6}$$

The standard deviation of the residuals depends on the residuals themselves as well as the degrees of freedom. The degrees of freedom is the difference between the number of observations and the number of statistics, or number of fields we are calculating. Since we have 2293 observations in the final dataset and 17 variables (including the bias), we have 2276 degrees of freedom. So the Standard deviation of the residuals, σ_r is

$$\sigma_r = \sqrt{\frac{\sum e^2}{n - 17}} \tag{7}$$

This is the typical residual distance from the predicted value to the observed value.

Results

We will see each weight coefficient, W in the below table. Also, Figure 2 shows each individual regression line through each of the continuous fields.

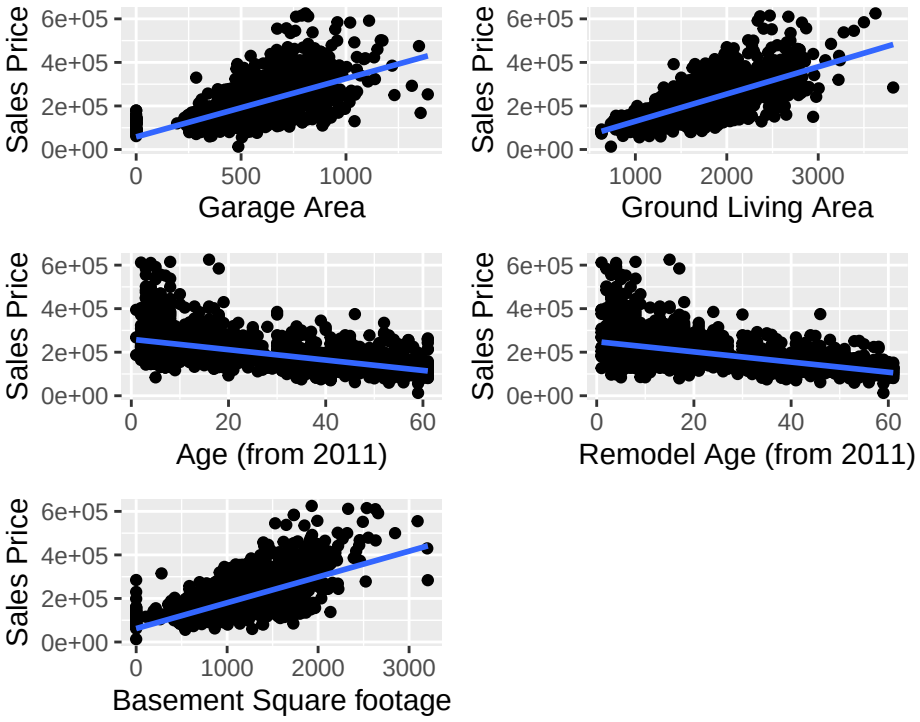

Figure 2: Sales

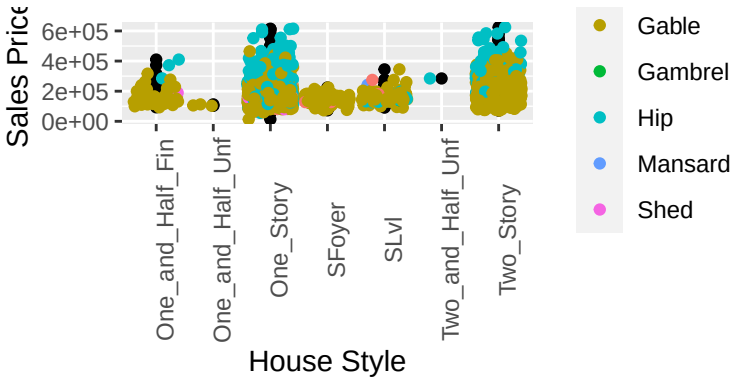

Figure 3: Sales vs Roof type

As we can see, it would appear no individual variable will accurately predict sales price thus the need for a multivariate linear regression model for variables *Ground Living Area (GrA)*, *Garage Area (GA)*, *Age (A)*, and *Remodeling Age (RA)*, *Total Basement Area (TBA)*, *Roof Style* ($1_A(r)$), and *House Style* ($1_A(h)$), where

$$1_{\text{Roof}}(x) = \begin{cases} 1 & x \in \text{Roof style} \\ 0 & \text{otherwise} \end{cases}$$

and

$$1_{\text{House}}(x) = \begin{cases} 1 & x \in \text{House style} \\ 0 & \text{otherwise} \end{cases}$$

.
Our model takes the form of

$$\text{Sales Price} = W_0 + W_1\text{GA} + W_2\text{GrA} + W_3\text{A} + W_4\text{RA} + W_5\text{TBA} + W_h 1_{\text{House}}(x) + W_r 1_{\text{Roof}}(x) \quad (8)$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20540.44190	10938.162388	1.8778695	0.0605268
Ground Area	78.78379	2.687397	29.3160241	0.0000000
Garage Area	66.46058	4.932884	13.4729645	0.0000000
Total Basement Area	47.26999	2.759660	17.1289197	0.0000000
Age	-713.84893	72.658045	-9.8247748	0.0000000
Remodel Age	-485.34974	70.547549	-6.8797534	0.0000000
(House Style)One and Half Unfinished	38676.71032	20929.223199	1.8479764	0.0647354
(House Style)One Story	10075.57272	5038.203586	1.9998344	0.0456368
(House Style)Foyer	21901.85237	6332.012148	3.4589088	0.0005523
(House Style)Split	14842.01515	5697.652556	2.6049351	0.0092490
(House Style)Two and Half Unfished	11840.94594	35561.992602	0.3329663	0.7391904
(House Style)Two Story	3287.78850	4988.523676	0.6590704	0.5099172
(Roof Style)Gable	-15103.72460	8204.094167	-1.8409984	0.0657519
(Roof Style)Gambrel	10987.29079	36285.061696	0.3028048	0.7620663
(Roof Style)Hip	5920.47842	8292.759401	0.7139335	0.4753415
(Roof Style)Mansard	-52977.77644	15618.923342	-3.3918968	0.0007060
(Roof Style)Shed	-22411.76604	17744.281491	-1.2630416	0.2067036

Calculated R squared value

```
## [1] 0.8018319
```

Residual Standard Error

```
## [1] 35179.49
```

Discussion

In order to answer how well the model performed, we discussed using the $\tilde{R}^2$ and Standard deviation of the residuals. The $\tilde{R}^2$ will give an idea of how much of the variance in Sales Price is accounted for by our model and the standard deviation of residuals will give us the typical distance between predicted and actual values.

We also see that the multivariate linear regression is a better predictor than using a single variable. The greatest R^2 value we get from any of the individual variables is 0.564 using *Ground Floor Area*. Using these 7 yielded better results.

R^2 Value

We want to see the predictive power of a multivariate linear model to predict Sales Prices of a house in Ames, Iowa. Just taking a cursory glance at Figure 2, no single variable is going to model sales price very well, hence the need for a multivariate approach. Using 6 continuous variables and 2 categorical variables, we get an $\tilde{R}^2$ value of 0.80, meaning 80 percent of the variance in Sales Price can be explained by our linear model. One may think is is a somewhat decent model, but let's look at the residual standard error.

Residual standard Error

Note that our residual standard error comes to approximately 35,100. Is this an acceptable average residual? It depends on the interpretation and the need. Note that the average sales price for the dataset is 194,718, with a min and max of 13,100 and 625,000, respectively. So our average error is approximately 18 percent of the average sales price. This is likely not going to be the best suited model for price prediction and perhaps adding more fields may yield better results. Additionally, we are making an assumption about linearity in the dataset.

References

De Cock, Dean. 2011. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project." *Journal of Statistics Education* 19 (3).

Mora-Garcia, Raul-Tomas, Maria-Francisca Cespedes-Lopez, and V. Raul Perez-Sanchez. 2022. "Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times." *Land* 11 (11). https://doi.org/10.3390/land11112100.

Olga, Bormpotsialou, and Rovolis Antonios. 2019. "Housing Construction as a Leading Economic Indicator." *Studies in Business and Economics* 14 (3): 33–49.

Truong, Quang, Minh Nguyen, Hy Dang, and Bo Mei. 2020. "Housing Price Prediction via Improved Machine Learning Techniques." *Procedia Computer Science* 174: 433–42. https://doi.org/https://doi.org/10.1016/j.procs.2020.06.111.

Zhang, Qingqi. 2021. "Housing Price Prediction Based on Multiple Linear Regression." *Scientific Programming* 2021 (October): 1–9. https://doi.org/10.1155/2021/7678931.

Appendix

This section will include code for building the dataset as well as some of the produced tables

Load Libraries

```
library(AmesHousing)
library(dplyr)
library(geodist) # Calculate Distance (geodesic)
library(tidyverse)
library(knitr) #For generating tables in the results section
library(xtable) #For generating tables in the results section
library(patchwork) # For creating the plots using ggplot
```

Clean data and create features

```
library(AmesHousing)
library(tidyverse)
library(patchwork)
library(ggplot2)
library(knitr) #For generating tables in the results section
library(xtable) #For generating tables in the results section

ames <- make_ordinal_ames()


### Get age from 2011
amesfixed <- ames %>%
  mutate(Age = 2011 - Year_Built)
### Get age of remodelling from 2011
amesfixed <- amesfixed %>%
  mutate(RemodelAge = 2011 - Year_Remod_Add)
### Housekeeping functions
# amesfixed <- rename(amesfixed, Lot.Area = `Lot Area`)
# amesfixed <- rename(amesfixed, Gr.Liv.Area=`Gr Liv Area`)
# amesfixed <- rename(amesfixed, Garage.Area=`Garage Area`)
# amesfixed <- rename(amesfixed, MS.Zoning=`MS Zoning`)
### Remove outliers as specificed in DeCook
amesUse <- amesfixed %>%
  filter(Gr_Liv_Area <= 4000 & Year_Built >= 1950)

amesUse <- amesUse %>%
  select(Gr_Liv_Area, Garage_Area,Total_Bsmt_SF, Age, RemodelAge, Sale_Price, House_Style, Roof_Style) %
  drop_na()

model <- lm(Sale_Price ~ Gr_Liv_Area + Garage_Area + Total_Bsmt_SF + Age + RemodelAge + factor(House_Sty
model %>%
  summary() %>%
  xtable() %>%
  kable()
### Get residual values
summa <- summary(model)
### Get R^2
summa$adj.r.squared
```

```
summa$sigma


# Or one may do this....
dof<-6 #Since there are 6 variables
sqrt(sum((summa$residuals)^2) / (length(summa$residuals) - dof))
```

Plot graphs

```
p1 <- ggplot(amesUse, aes(x=Garage_Area, y=Sale_Price)) +
  xlab("Garage Area") +
  ylab("Sales Price") +
  geom_point() +
  geom_smooth(method=lm)

p2 <- ggplot(amesUse, aes(x=Gr_Liv_Area, y= Sale_Price)) +
  xlab("Ground Living Area") +
  ylab("Sales Price") +
    geom_point() +
  geom_smooth(method=lm)

p3 <- ggplot(amesUse, aes(x=Age, y= Sale_Price)) +
  xlab("Age (from 2011)")+
  ylab("Sales Price") +
    geom_point() +
  geom_smooth(method=lm)

p4 <- ggplot(amesUse, aes(x=RemodelAge, y= Sale_Price)) +
  xlab("Remodel Age (from 2011)")+
  ylab("Sales Price") +
    geom_point() +
  geom_smooth(method=lm)

p5 <- ggplot(amesUse, aes(x=Total_Bsmt_SF, y= Sale_Price)) +
  xlab("Basement Square footage")+
  ylab("Sales Price") +
  geom_point() +
  geom_smooth(method=lm)


p1  + p2 + p3 + p4  + p5 + plot_layout(ncol = 2)
```