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

As stated before, we will be using a small portion of the Ames Housing Dataset. We will remove the majority of the data and focus on the fields that are easily calculated or already provided in the dataset. The fields chosen are *Ground Living Area*, *Garage Area*, *age (from 2011)*, *age of remodeling (from 2011)*, and *Distance to school*. These fields are easily retrieved from most assessor data and are easy to understand. We also wanted to include data that is possibly going to be important to home buyers. While the style of the house may vary, living space, garage space, age, and distance nearest school seem to be reasonable indicators for home values.

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.

Simple Linear Regression

As stated in the Introduction, 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:

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

Since we are using linear regression, we can analytically find the best fit weights $W_0 \dots W_n$ using the following formula:

$$W = (X^t X)^{-1} X^t Y \tag{4}$$

Where W is the $1x(n+1)$ weight matrix.

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 R^2 statistic. R^2 is defined as the amount of variance of observations that is accounted for by the linear regression model. It is calculated as follows:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{5}$$

The closer $\frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$ is to 0, the greater the predicting power of the model. This should make sense intuitively, since if $(y_i - \hat{y}_i)^2 \approx 0$ then the closer the predicted and the observed values are.

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 residualse 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 2912 of observations in the final dataset and 6 variables (including the bias), we have 2906 degrees of freedom. So the Standard deviation of the residuals, σ_r is

$$\sigma_r = \sqrt{\frac{\sum e^2}{n-6}} \tag{7}$$

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 1 shows each individual regression line through each of the pertinent fields.

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)*, and *Distance to school (D)*. Our model takes the form of

$$\text{Sales Price} = W_0 + W_1\text{GA} + W_2\text{GrA} + W_3\text{A} + W_4\text{RA} + W_5\text{D} \tag{8}$$

	Estimate	Standard Error	t value	Pr(>|t|)
(Intercept)	45611.47362	3387.018207	13.466557	0
GrA	74.46729	1.781833	41.792518	0
GA	87.81362	4.272606	20.552705	0
A	-650.49109	32.520872	-20.002265	0
RA	-423.87212	45.662808	-9.282656	0
D	15670.59440	1011.644060	15.490225	0

Calculated R squared value

```
## [1] 0.7524802
```

Residual Standard Error

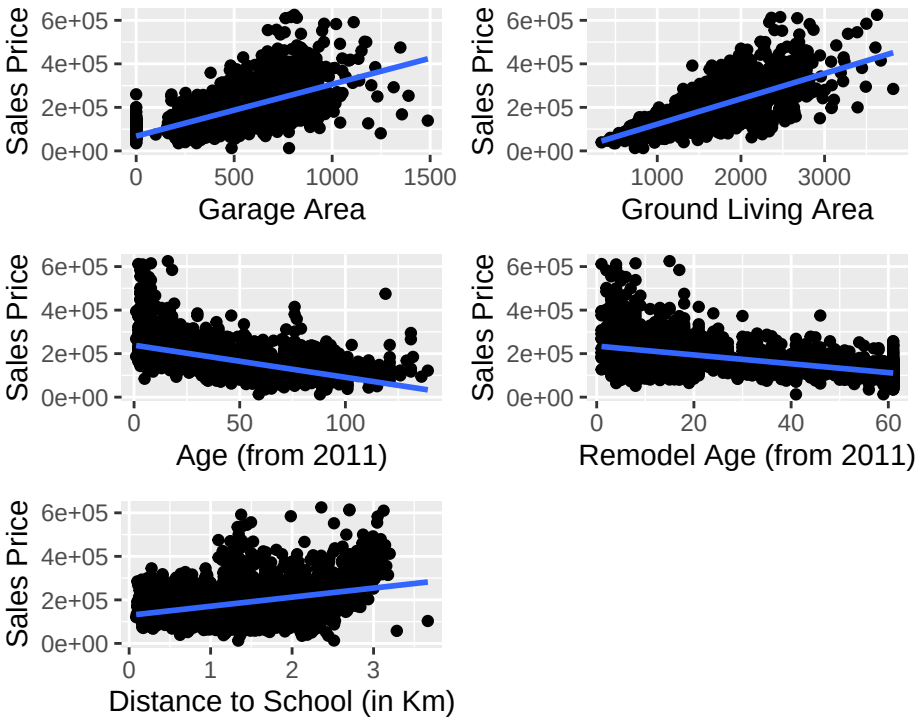

Figure 1: Sales Price vs Variable plots

```
## [1] 39098.66
```

The final equation for predicted Sales Price is thus

$$\text{Sales Price} = 45611.47362 + 74.46729 * \text{GrA} + 87.81362 * \text{GA} + -650.49109 * \text{Age} + -423.87212 * \text{RA} + 15670.59440 * \text{D}$$

Discussion

In order to answer how well the model performed, we discussed using the R^2 and Standard deviation of the residuals. The R^2 gives 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.

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 1, no single variable is going to model sales price very well, hence the need for a multivariate approach. Using 5 variables, we get an R^2 value of 0.75, meaning 75 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 39,100. Is this an acceptable average residual? It depends on the interpretation and the need. Note that the average sales price for the dataset is 180,796.10, with a min and max of 12,789 and 755,000, respectively. So our average error is approximately 21.6 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. This may not actually be the case and in fact a different model may be better suited for modeling sales prices.

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

```
### Load data
data("ames_raw")
data("ames_geo")
data("ames_schools_geo")

### First calculate the distance to the closest school (in Kilometers)

dist <- geodist_vec(x1=ames_schools_geo$Longitude, y1= ames_schools_geo$Latitude,
                    x2=ames_geo$Longitude, y2=ames_geo$Latitude,
                    measure = "geodesic") / 1000

### get minimum distance
mdist <- apply(dist, 2, min)
ames_geo$Distance <- mdist

### Join the tables based on Parcel ID
jnd <- ames_geo %>%
  select("PID","Distance")

ames_raw <- merge(ames_raw, jnd, "PID")

### Get age from 2011
amesfixed <- ames_raw %>%
  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)
```

```
### Select pertinent fields and drop NAs
amesUse <- amesUse %>%
  select(Gr.Liv.Area, Garage.Area, Age, RemodelAge, SalePrice, Distance) %>%
  drop_na()
```

Create linear regression model and output table

```
### Create the model and create table in results section
model <- lm(SalePrice ~ Gr.Liv.Area + Garage.Area + Age + RemodelAge + Distance, amesUse)

model %>%
  summary() %>%
  xtable() %>%
  kable()

### Get residual values
summa <- summary(model)

### Get R^2
summa$r.squared

### Get standard deviation of residuals

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=SalePrice)) +
  xlab("Garage Area") +
  ylab("Sales Price") +
  geom_point() +
  geom_smooth(method=lm)

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

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

p4 <- ggplot(amesUse, aes(x=RemodelAge, y= SalePrice)) +
  xlab("Remodel Age (from 2011)")+
  ylab("Sales Price") +
```

```
    geom_point() +
  geom_smooth(method=lm)

p5 <- ggplot(amesUse, aes(x=Distance, y= SalePrice)) +
  xlab("Distance to School (in Km)")+
  ylab("Sales Price") +
  geom_point() +
  geom_smooth(method=lm)


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