Housing price prediction based on Multiple linear regression and PCR

Group members:

Chengyan(Kurt) Ji
Rong Li
Yizhe He

Roles:

Data processing: Chengyan(Kurt) Ji

Part 1&2: Chengyan(Kurt) Ji

Part 3.1 model and coding: Rong Li

Part 3.2 model and coding: Yizhe He

Part 4: all of us

Paper formatting: Rong Li

1 Introduction

1.1 Background

Housing price has been a very important subject of interest as it plays an important role in both people’s daily life and the investment market. As housing prices usually have a similar pattern compared to the overall performance of the economy, people in the financial industry are still able to come up with generalized models that can describe the correlation between the price and observable factors.

1.2 Current Models

Current models employed by professionals on estimating house price include income approach, cost approach, comparative approach, and hedonic price approach. These models have different preferences over the choice of variables and coefficients. These choices include weighted average models, neural networks, and regression models.

- In general, the income approach uses the cash flow in the real estate market as its variable of interest, but this method is both costly in terms of monetary and timely spendings.
- The cost approach focuses on the cost for the real estate developer to build the house. This method is bad at taking into account the fact that the real estate market is not an efficient market hence the market performance of some houses is neglected, as their prices depend more on the marketability rather than just on the cost.
- The comparative approach takes insights from similarities in the housing market as it takes information from a similar unit on the market and the model is implemented with a correction coefficient from the professionals in the real estate industry. The disadvantage is that it is heavily affected by subjectivity.
- The hedonic approach is the quantitative method of the comparative approach. It is run on a regression model so it does not have the human bias in selecting coefficients. It is cost-effective so it is massively used by the industry and the academies. Neural network methods are also employed in the hedonic approach. The general disadvantage is that it lacks persuasiveness to the client because it does not include a comparative case from the real market
- Other than models included above, it is plausible that we can use other relevant information to predict the price of a certain house. The intuition is that the location of the house, the historical information about the house and the surrounding infrastructure all have an impact on the market price of the house. Since the demand and supply in the housing market is heavily affected by many daily life factors, we see that if we could collect as much relevant information as possible, we will have a great chance of knowing how the demand and supply are going to look like for a particular housing unit, hence predicting its performance in the housing market. Therefore, given enough information, we may very likely find the empirical pattern between these factors and the price, which we can use to model their actual causal effects on the price and predict the price based on their values.

1.3 Our purpose

The purpose of this paper is to demonstrate the methods used in evaluating house pricing based on geolocational data and other informational data such as age, nearby convenience stores and transaction date. The models employed aim to best describe the relationship between the factors and the price with minimum error.

2 Data

2.1 Data source and overview

The data is collected from UCI machine learning repository. It is contributed by professor I-Cheng Yeh from the Department of Civil Engineering at Tamkang University in Taiwan.

The data set is the market historical data of real estate valuation, collected from Sindian Dist., New Taipei City, Taiwan. There are mainly three types of variables included for the modeling. The first type of data is the geolocational data, which includes the coordinates. The second type is time data, including the transaction date and the age of the house. The last type is other relevant information that describes the neighborhood effects, including the number of convenient stores nearby and the distance to the nearest MRT station.

The variable included in the models are as follows:

1. X1 = the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
2. X2 = the month of the transaction date
3. X3 = the house age (unit: year)
4. X4 = the distance to the nearest MRT station (unit: meter)
5. X5 = the number of convenience stores in the living circle on foot (integer)
6. X6 = the geographic coordinate, latitude. (unit: degree)
7. X7 = the geographic coordinate, longitude. (unit: degree)
8. Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

2.2 Data Description

For the convenience for modeling, the data set is split randomly into two portions, where one contains 70% of the data and is the training set; the other one containing 30% will be used for testing the models

Among the 7 independent variables, the month in the transaction date can be regarded as a categorical variable, while the rest can be regarded as numeric. Before models are run on these data, we show an exploratory analysis where we can have some insight on how the variables are related to each other.

The first thing to check is whether there exists collinearity within the dataset. This problem can violate the basic linear model assumptions and can cause serious problems. With the correlation plot, we see that there are some variables that are correlated (Fig. 1). Since we are only looking at the housing market within a small area, the coordinates are likely correlated to the distance to the train station. However, the correlation between latitude and longitude may just be random. Later in the report, this issue will be addressed with numeric analysis.

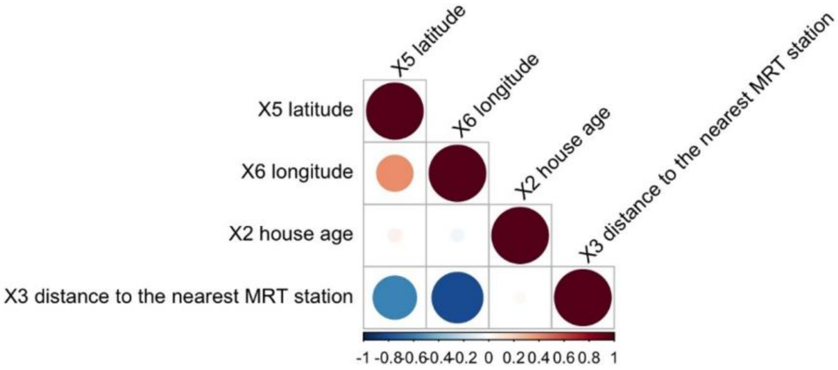


Fig. 1. The correlation plot shows a mild correlation between X6 longitude and X5 latitude, as well as between X5 latitude and X3 distance to nearest MRT station, and X6 longitude and X3 distance to nearest MRT station

If we solely look at the house prices, we can see a regular pattern with most of the houses being priced between 250000 to 500000 dollars a Ping (Fig. 2). The median is around 450000 dollars a Ping. There are some extremely expensive houses that are dragging the distribution to become right-skewed.

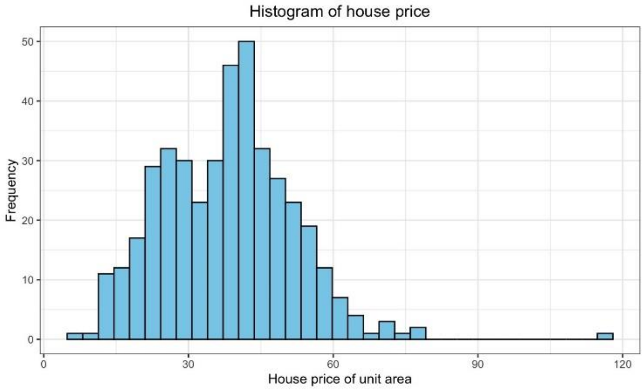


Fig. 2. The histogram of housing price shows a right skewed distribution, with the median at around 450000 and the mode of around 450000.

What's more, if we plot the houses with respect to the time they were constructed, we get a time series data (Fig. 3). From the plot we can see regardless of the time the price maintains at a steady level. We can make some inference that maybe time is not affecting the price in a detectable way. However this can be addressed by the fact that the data set only contains barely over 1 years of data.

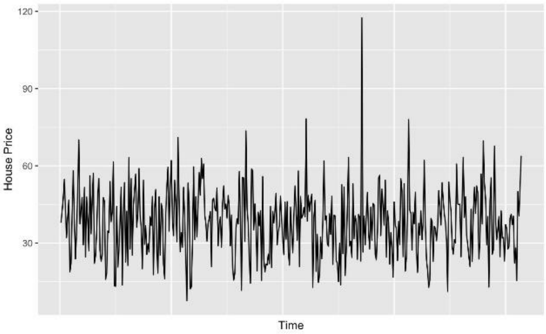


Fig. 3. The time series is the price of each unit versus the time lapsed since construction. We can see that out of the 414

units constructed, there are very few units that are significantly different from others. So the price has not changed a lot across time.

Another variable worth looking at is the age of the house, which may play an important part in the price determination, intuitively. When we plot the ages of each unit, we see that the ages have a very large range, from brand-new to over 40 years old (Fig. 4). This shows that the area we are looking at has been historically a site for housing. Over half of the houses are between 10 and 20 years old, and with an increasing number of new houses built, we can see a growing market from this area.

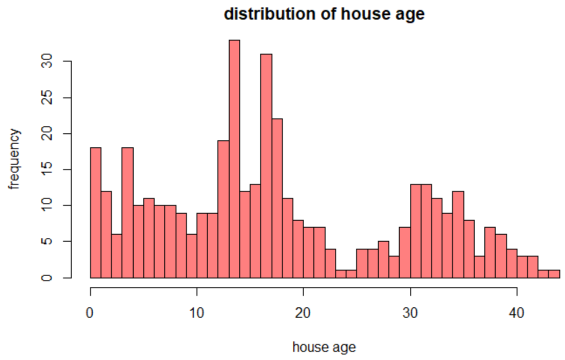


Fig. 4. The distribution of house age

The distribution shows that out of the 414 units, the age distributes mostly evenly but with several extreme cases like the number of units at 15 and 18 years of age is very high and the number of units at 25 years of age is very low. We can see the house construction has remained stable, but has seen some soaring and low tides across time. We can also see that recent house construction has an increasing trend.

If we separate the prices by the month they are built, we may see some more interesting patterns (Fig. 5). October has the lowest mean price among all months, while March has the highest. We can also see that July has a very wide range of transaction amounts, with the majority ranging from 250000 a Ping to 450000 a Ping, and the other months have smaller ranges.

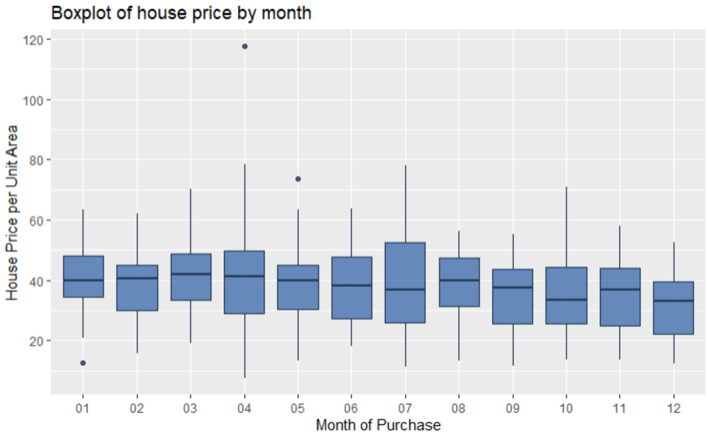


Fig. 5. Boxplot shows the price in each month

When we look at how location affects housing prices, we find it surprising that both longitude and latitude have a positive effect on the price (Fig 6 and 7). While this is somehow counter-intuitive, we may go back to something we talked about earlier. From the correlation plot (Fig. 1), we find that as the longitude and the latitude increase, the distance to the nearest MRT station decreases. This might be due to the specific data set we are looking at, but as we plot the price against the distance to MRT station, we find a very intuitive pattern (Fig. 8): the price is decreasing with respect to the distance. This explains the visible correlation between the coordinates and the housing price, and this specific pattern will play an important part in the model analysis later.

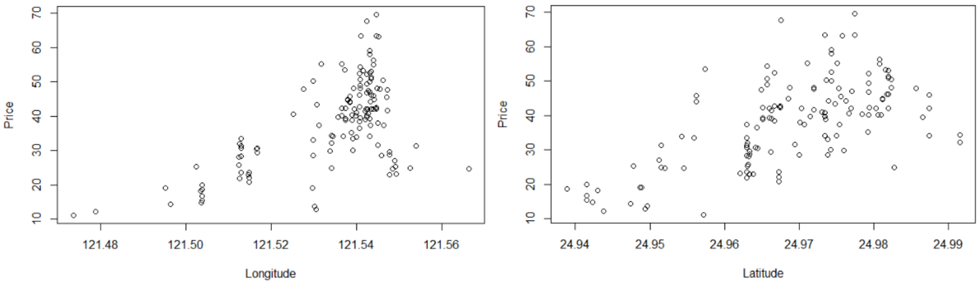


Fig. 6 & **Fig. 7.** Scatter plots of price against longitude and latitude show positive correlation

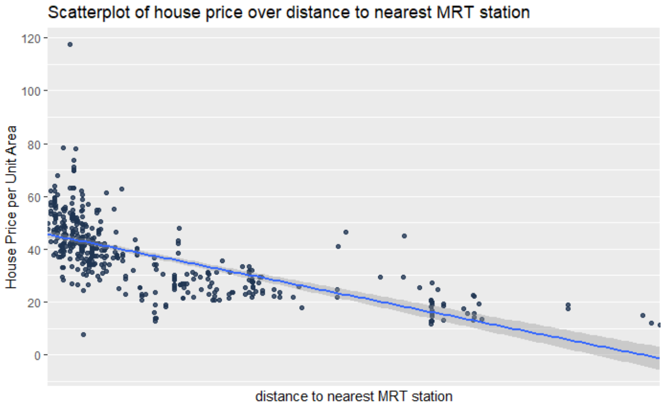


Fig. 8. Scatter plot of price against distance to nearest MRT station shows negative effect from the distance on the price

3 Methodology

3.1 Multiple Linear Regression

Multiple linear regression is commonly used to detect relationships among variables. Main model assumptions of it are:

1. The mean function is the linear combination of predictors
2. Errors are uncorrelated with 0 mean and constant variance of σ^2
3. Errors form to normal distribution

Hence, after fitting the model, it is necessary to test whether these assumptions hold true. And the design matrix in regression is made from the observations, we must be careful if the design matrix has rank deficiency, namely, there is collinearity among predictors. After fitting the model, we can apply several partial F-tests or permutation tests to check the significance of each predictor.

3.1.1 Fit the model

We first came up with multilinear regression. First, we divide the observations into training sets and testing sets according to the ratio of 7:3. We regard months as categorical data and others as numeric data. The box plot shown below indicates the effects of month on price.

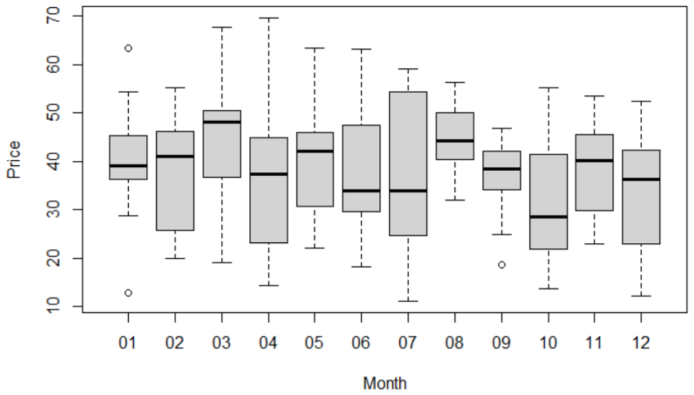


Fig. 9. Box plot of Price ~ Month

We notice that, after taking the logarithm of "distance to the nearest MRT station", it is easier to find the linear relationship between price and distance.

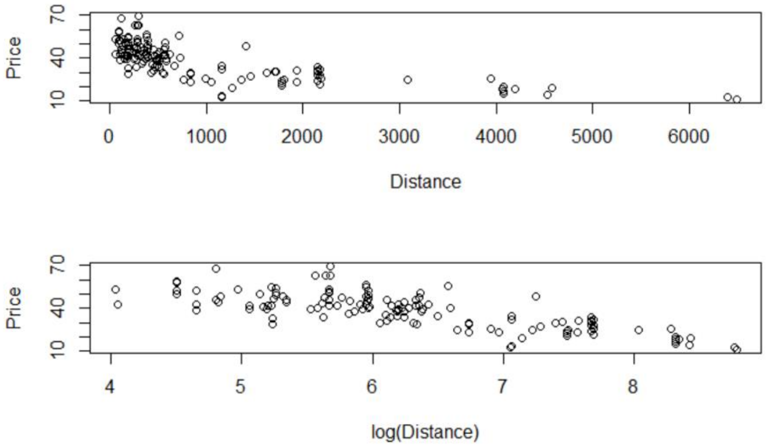


Fig. 10. Scatter plot before and after taking logarithm on Distance

Hence, it is not unreasonable to take the logarithm of "distance to the nearest MRT station". After doing this, we start fitting models.

Initially, we include all the 7 variables —— transaction date, the month of the transaction, the house age, the distance to the nearest MRT station, the number of convenience stores in the living circle on foot, latitude, longitude. With all the main predictors and interaction parts with the month included, namely Price~Date+Month+House_age+log(Distance)+Stores+Latitude+Longitude+House_age:Month+log(Distance):Month+Stores:Month+Latitude:Month+Longitude:Month.

We got a model with 0.8661 R-square, however, almost all the predictors are not significant. And an analysis of covariance tells us that among all the interaction parts, only "Stores : Month" has effects on price.

Secondly, we remove all the interaction parts except for "Stores : Month" and refit the model, namely Price~Date+Month+House_age+log(Distance)+Stores+Latitude+Longitude+Stores:Month. We get R-square equals to 0.7836 and some of the predictors are significant. Whereas, the p-value of "Date" is 0.8285, which is not significant.

Thirdly, we refit the model after removing "Date", namely Price ~ Month + House_age + log(Distance) + Stores + Latitude + Longitude + Stores:Month, which turns out to have 0.7836 of R-square.

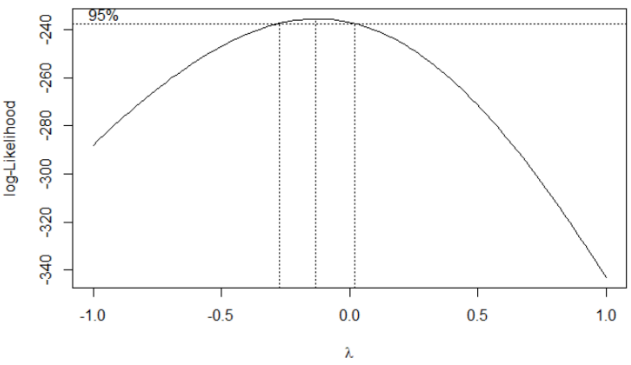


Fig. 11. The Box-Cox Test result shows 0 is included in the confidence interval

From the Cox-Box Test, 0 is included in the confidence interval. Therefore, we have to take the logarithm of the response.

Finally, we take the logarithm of the response("Price") from the result of the Box-Cox Test and refit the model, namely, log(Price) ~ Month + House_age + log(Distance) + Stores + Latitude + Longitude + Stores:Month. Fortunately, we got 0.8473 of R-square and all the coefficients of numeric predictors are significant.

3.1.2 Model diagnostics and detection of unusual observations

- Collinearity detection

We use VIF to check whether there is collinearity or not.

Table 1

VIF of numeric predictors

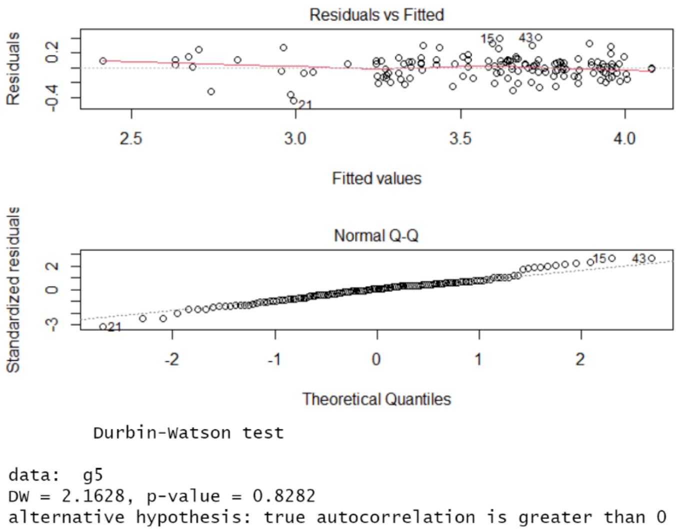
predictors	House_age	log(Distance)	Stores	Latitude	Longitude
VIF	1.02	2.62	1.96	1.53	1.91

With all the VIF values less than 10, we conclude that there is no collinearity between variables.

- Detect unusual observations

Firstly, we apply Cook's distance to identify high influential points, all the Cook's distances are less than 1, so there is no high influential point. Secondly, we use the hat matrix to detect high leverage points. Specifically, high leverage points are points with $h_i > 2\overline{h_i}$. And we found 6 points. At last, t-distribution is adopted in finding outliers. Compared with 0.025 quantile after Bonferroni correction, no outliers exist. In short, there are only 6 high leverage points.

- Test model assumptions



```
        Durbin-Watson test

data:  g5
DW = 2.1628, p-value = 0.8282
alternative hypothesis: true autocorrelation is greater than 0
```

Fig. 12. Residual plots and Q-Q plot and DW-test

Firstly, we deem there is constant variance from residual plot with a horizontal line. Secondly, the quantile-quantile plot tells us that we cannot reject the assumption of normality. Next, the Durbin-Watson test returns a p-value of 0.8282, which shows that true autocorrelation is equal to 0.

- Refit the model after removing unusual observations

From the model diagnostics and model checking, we should refit the model after removing unusual observations (high leverage points).

However, we got R-squared only 0.8383, which is smaller than when we don’t remove high leverage points(0.8473). This is possible because there are good high leverage points and bad high leverage points. Good high leverage points go with the trend and only bad high leverage points go against the trend. What is more, AIC for high-leverage-removed model and not-removed model are -444.2358 and -471.6476 respectively. Hence, we should keep the 6 high leverage points.

- conclusion

In conclusion, the final regression model we constructed is log(Price) ~ Month + House_age + log(Distance) + Stores + Latitude + Longitude + Stores:Month and without removing high leverage points.

3.1.3 Prediction

The following table summarizes the first 5 observations in the test set, we can see how well the model is predicting data.

Table 2

Difference between predicted value and true value of the first 5 values

Predicted value	50.16083	45.33749	50.18856	41.43325	35.65820
True value	37.9	47.3	54.8	43.1	40.3

We calculate the Root Mean of Squared Error using $RMSE = \sqrt{\frac{1}{N}\sum(ture\ value - predicted\ value)^2}$

Where N is the number of rows of test set. And we got the value of 9.859559.

3.2 Principal Component Regression

In this section we have two steps. First is to perform the principal component analysis (PCA), and then fit the principal component regression (PCR) model.

3.2.1 PCA

Principal component analysis is commonly used for dimension reduction in high dimensionality data by projecting each data point onto only a few linear combinations of the original variables called "principal components (PCs)" to obtain lower-dimensional data while preserving as much of the data's variation as possible.

The idea of principal component analysis is based on spectral decomposition of the covariance matrix of the dataset. The geometric meaning of PCA is to rotate all the data points onto new axes such that the direction maximizes the variation of data.

The interpretation of PCA can be challenging since it projects all the data points onto new axes. However, there are some nice properties of this method:

- The PCs are uncorrelated, meaning that their correlation coefficients are 0. Geometrically, they are orthogonal to each other.
- The variances of the PCs are in decreasing order.
- The principal components analysis preserves all the information from the original data.

Table 3

Cumulative variance for principal components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.0619	0.7229	0.3671	0.2546	0.0482	0.0000	0.0000
Cumulative variance	0.5915	0.8656	0.9363	0.9702	0.9988	0.9999	1.0000

Note: The data is being standardized before performing PCA

The above chart shows the results after we perform PCA on the training data set. Since there are 7 predictors, we got 7 principal components, but we don't really need all of them. Why? Because we noticed that the cumulative variance of the first three PCs already have 93% of the original, and the rest only have less than 7%. So we can say most of the total variation of the original data is explained by the first three PCs. Here the idea of dimension reduction comes in, we think the first three PCs are "important" for interpretation, while the remaining PCs are merely "noise".

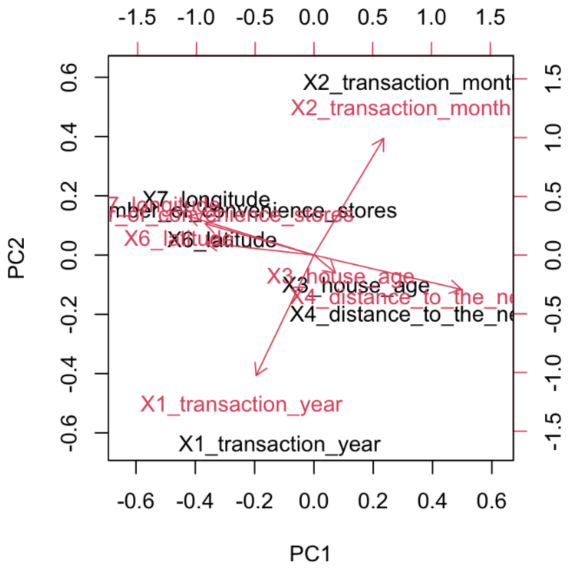


Fig. 13. The biplot of PC1 vs. PC2

Figure 13 is the biplot of PC1 vs. PC2. This plot can help us to visualize the PCA and easily interpret the results. We can see that the PC2 is heavily load on year and month, and PC1 is heavily load on the rest. In general, the method of dimension reduction can help us to better understand the overall structure of our data especially on high dimensions.

3.2.2 PCR

After briefly introducing the concept of principal component analysis, we will fit what's known as the principal component regression (PCR) model in this section. Obviously, PCR is closely related to PCA. Basically, we get a "new" set of predictors after we perform PCA. In fact, this new set of predictors are a linear combination of the original predictors. Thus, we can use our new predictors to predict the response variable.

In practice, we only focus on the prediction when we use PCR. There is one thing we need to think about carefully: How many PCs should we use? In the previous section, we found that the first three PCs explained around 90% variation of the data set. So one reasonable choice is to use three PCs. Since our main goal is to make predictions on house prices, we can split the data into training sets and testing sets. On the training data, we can try to fit different models using different numbers of PCs as our tuning parameter. Then, we can make a prediction on the testing data and calculate their corresponding prediction root mean squared error (RMSE).

Table 4

The RMSE for each PCR model

Number of PCs	1	2	3	4	5	6	7
RMSE	10.18612	10.11354	9.85296	9.67931	9.61739	9.60733	9.44820

Here is the summary table for each model and their prediction RMSE in testing data. It looks like when we use 7 PCs, the RMSE is the lowest. One thing to notice is that when the number of PCs is 5 or more, the RMSE decreases very slightly. In order to balance the purpose of dimension reduction and the goal of minimizing the RMSE, we could choose the model with 4 PCs or 5 PCs.

Another way to help us select the number of PCs is cross-validation (CV). In general, cross-validation is a technique for assessing how the results of a statistical model will generalize to an independent data set. Cross-validation is a resampling method that uses different portions of the data to train and test a model on different iterations. The common way to do cross-validation include 10-fold CV, leave one out CV (LOOCV). We will perform 10-fold cross-validation in this report.

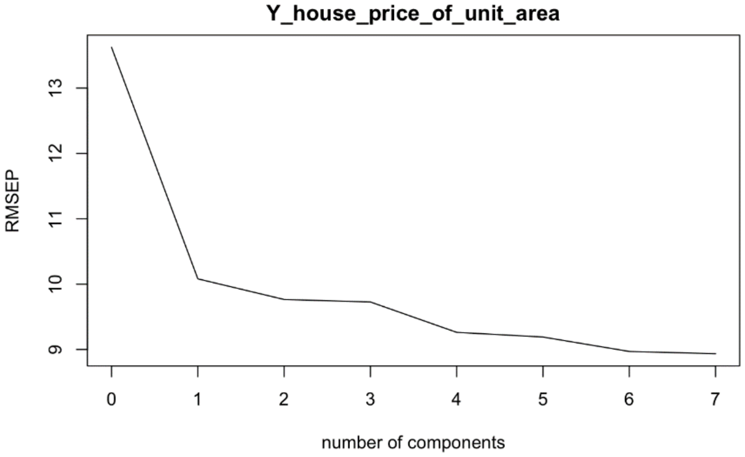


Fig. 14. The scree plot for PCR

This scree plot shows the RMSE results for each PCR model by 10-fold CV. We can see that the trend of RMSE is decreasing. This implies the more PCs we use, the less RMSE we would have. We could use 7 PCs since it gives us the smallest RMSE or we could use 4 PCs since it balances the dimensionality and prediction error.

4 Model performance comparison

4.1 Summary and discussion

- Multiple Linear Regression

The multiple linear regression focuses on the correlation between the predictors and the response variable. It involves multiple tests on the assumptions and on the predictors in order to achieve the best fit. In our model, we removed some interaction terms between the categorical variable and other variables after several t-tests. Then we dropped the ‘Date’ variable and used log transformation after the Box-cox test. After testing model assumption and removing unusual observations, the final model is as follows:

log(Price) ~ Month + House_age + log(Distance) + Stores + Latitude + Longitude + Stores:Month

The model is easy to read because every coefficient represents the marginal effect of the corresponding predictor to the response variable. We can get a sense of how each of the predictors is causing the price to change, including the direction and the magnitude. However, linear regression relies on assumptions about the model and any violation may harm the validity of the prediction. In our model, we dropped multiple regressors and applied transformation in order to avoid such violations. On the other hand, the variances of the coefficients are usually larger which can cause instability of the model.

- Principal Component Regression

For the Principal Component Regression, we decided to use 4 PCs as our final model since it takes into account dimension reduction and prediction accuracy at same time. In general, PCR is a model with strong applicability. It can help us to better understand the structure of high dimensional data. Compared to the MLR, there is no strict assumption in PCR. Therefore, it is more flexible than MLR and makes the diagnostic procedure much easier. It can address the issues of multicollinearity and over-fitting when MLR is often encountered. However, there are also some limitations for this method:
the interpretability for PCR is very vague, sometimes, there is even no actual meaning for PC at all. Although PCs try to capture the maximum variation among the features in data, if we don’t select the number of PCs with care, it may lose some information as compared to the original data.

- Impact

Our group come up with 2 easy-to-apply models predicting house price in New Taipei City, Taiwan, which can provide pricing guidance for real estate developers and provide homebuyers with a reasonable predicted purchasing price.

4.2 Final recommendations

We recommend principal components analysis model. Since PCR has several advantages compared to ordinary multiple linear regression.

- MLR is parametric regression, which means it has a strong model assumption. We must deal with the model diagnostic after we fit the model such as detecting outliers.
- PCR are much more flexible than MLR. We can really just ignore any assumption we made in MLR.
- RMSE of MLR and PCA are 9.86 and 9.68 respectively, namely PCA has a better prediction result.