Linear Regression Analysis for the Kings County's (Seattle, WA) House Market.

by Y. Kostrov

Contents

Overview

The purpose of this project is to analyze a data set containing data about houses sold in Kings County (Seattle, WA).

Overview

During the analysis:

Overview

During the analysis:

1. I will perform necessary data wrangling first.

Overview

During the analysis:

1. I will perform necessary data wrangling first.
2. I will build a Linear Regression Model with one explanatory variable.:

Overview

During the analysis:

1. I will perform necessary data wrangling first.
2. I will build a Linear Regression Model with one explanatory variable.:
   - I will check statistical assumptions for the linear regression model
   - I will explain the model, including intercept, coefficient for the explanatory variable, R^2, and ANOVA

Overview

During the analysis:

1. I will perform necessary data wrangling first.
2. I will build a Linear Regression Model with one explanatory variable.:
    - I will check statistical assumptions for the linear regression model
    - I will explain the model, including intercept, coefficient for the explanatory variable, R^2, and ANOVA
3. I will build a Multiple Linear Regression Model with many explanatory variables.

Overview

During the analysis:

1. I will perform necessary data wrangling first.
2. I will build a Linear Regression Model with one explanatory variable.:
   - I will check statistical assumptions for the linear regression model
   - I will explain the model, including intercept, coefficient for the explanatory variable, R^2, and ANOVA
3. I will build a Multiple Linear Regression Model with many explanatory variables.
   - I will check statistical assumptions for the multiple linear regression model
   - I will explain the model, including intercept, coefficients for the explanatory variables, R^2, and ANOVA

Business Problem

- The fair price of the house is a hard quantity to assess.

Business Problem

- The fair price of the house is a hard quantity to assess.
- Both sellers and buyers would like to know the best price for the house.

Business Problem

- The fair price of the house is a hard quantity to assess.
- Both sellers and buyers would like to know the best price for the house.
- Which features of the property would be the best predictors of the value?

Business Problem

- The fair price of the house is a hard quantity to assess.
- Both sellers and buyers would like to know the best price for the house.
- Which features of the property would be the best predictors of the value?
- I will build a regression model that helps predict the value of the house.

Business Problem

- The fair price of the house is a hard quantity to assess.
- Both sellers and buyers would like to know the best price for the house.
- Which features of the property would be the best predictors of the value?
- I will build a regression model that helps predict the value of the house.
- I will, also, check the necessary statistical assumptions for the regression model and explain the model's parameters.

Data Description

- The file called "kc_house_data.csv" in the data folder of the project holds the data for this project.

Data Description

- The file called "kc_house_data.csv" in the data folder of the project holds the data for this project.
- This project will use this data about Kings County's(Seattle, WA) housing market to create Linear Regression Model.

Data Description

- The file called "kc_house_data.csv" in the data folder of the project holds the data for this project.
- This project will use this data about Kings County's(Seattle, WA) housing market to create Linear Regression Model.
- The data file contains numerous columns with information about properties sold such as price, size of the living area, size of the basement, number of bedrooms, etc.

My Python Package

- While working on this project, I have created my own Python package with helping functions.

My Python Package

- While working on this project, I have created my own Python package with helping functions.
- The most important function in this package is "evaluate_model.py" (in the "src" folder). This function:

My Python Package

- While working on this project, I have created my own Python package with helping functions.
- The most important function in this package is "evaluate_model.py"(in the "src" folder). This function:
    - creates the model from the data frame.

My Python Package

- While working on this project, I have created my own Python package with helping functions.
- The most important function in this package is "evaluate_model.py" (in the "src" folder). This function:
    - creates the model from the data frame.
    - prints out the model summary of Linear Regression.

My Python Package

- While working on this project, I have created my own Python package with helping functions.
- The most important function in this package is "evaluate_model.py" (in the "src" folder). This function:
  - creates the model from the data frame.
  - prints out the model summary of Linear Regression.
  - performs the checks for the statistical assumptions of the Linear Regression.

My Python Package

- While working on this project, I have created my own Python package with helping functions.
- The most important function in this package is "evaluate_model.py" (in the "src" folder). This function:
  - creates the model from the data frame.
  - prints out the model summary of Linear Regression.
  - performs the checks for the statistical assumptions of the Linear Regression.
  - performs a lot of different visualizations.

Modeling

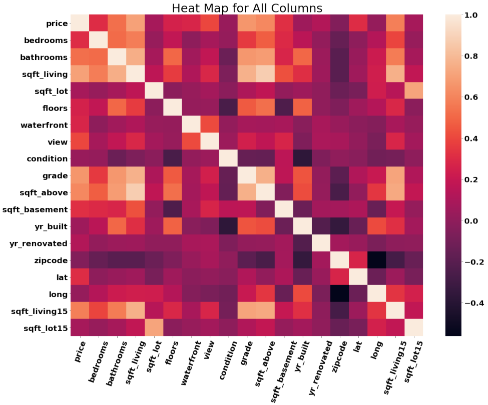
- My first goal was to create linear regression model with one independent variable.

Modeling

- My first goal was to create linear regression model with one independent variable.
- I created the correlation matrix and heat map for visualization purpose.

Modeling

- My first goal was to create linear regression model with one independent variable.
- I created the correlation matrix and heat map for visualization purpose.



Selection of the explanatory variable for the Linear Model

- "sqft_living" has the highest correlation of 0.71 with the "price".

Selection of the explanatory variable for the Linear Model

- "sqft_living" has the highest correlation of 0.71 with the "price".
- I build a regression model for the "price" to be predicted by "sqft_living".

Selection of the explanatory variable for the Linear Model

- “sqft_living” has the highest correlation of 0.71 with the “price”.
- I build a regression model for the “price” to be predicted by “sqft_living”.
- The model is $\ln(price) = 11.9524 + 0.0029 \cdot sqft\_living^{0.78}$

Checking Statistical Hypotheses:

This Linear Model satisfies all statistical assumptions of the Linear Regression, namely:

- Linearity.
- Normality.
- There is no heteroscedasticity present in the model.

Intercept and slope

Our model is

$$\ln(price) = 11.9524 + 0.0029 \cdot sqft\_living^{0.78}$$

- The model has the sample intercept of 11.9524 and the slope of 0.0029.
- To interpret the slope, we have to transform $\hat{x}$ and $\hat{y}$ towards original *sqft_living* and *price*.

Intercept and slope

Our model is

$$\ln(price) = 11.9524 + 0.0029 \cdot sqft\_living^{0.78}$$

- The model has the sample intercept of 11.9524 and the slope of 0.0029.
- To interpret the slope, we have to transform $\hat{x}$ and $\hat{y}$ towards original *sqft_living* and *price*.

To understand the change in *price* in percents we will use the following formula:

$$Gab100 \times \left[e^{0.0029\left(x_2^{0.78} - x_1^{0.78}\right)} - 1 \right]$$

Example

For example, if the *sqft_living* is $1000ft$ and we increase it to $1100ft$, we will get the change in price of

$$100 \times \left[e^{0.0029(1100^{0.78} - 1000^{0.78})} - 1\right] \approx 5.02\%.$$

Example

For example, if the *sqft_living* is $1000ft$ and we increase it to $1100ft$, we will get the change in price of

$$100 \times \left[e^{0.0029(1100^{0.78}-1000^{0.78})} - 1\right] \approx 5.02\%.$$

In this particular example, 10% change in sqft_living starting from $x_1 = 1000ft$ forces 5.02% change in price.

R^2

- The model has $R^2 \approx 0.45$.

R^2

- The model has $R^2 \approx 0.45$.
- This means that our model explains about 45% of the variation by using *sqft_living* as independent variable.

ANOVA

Is our model with one explanatory variable better than the model with zero explanatory variables?

- Our p-value for this model is $p = 0.000 < 0.05 = \alpha$.
- Since our p-value is 0, there is a 0% probability that the improvements that we are seeing with our one independent variable model are due to random chance alone.

col

Multiple Linear Model

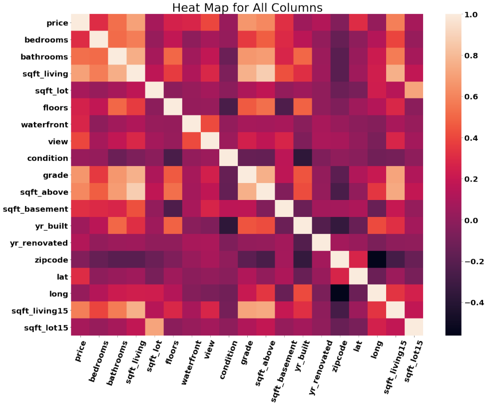
Now, I will build a Multiple Regression Model.

The goals for Multiple Linear Model:

- I want to improve R^2.
- I want to use more than one explanatory variable.

Choice of Explanatory Variables

I will use the highly correlated with the price features from the correlation matrix for the Multiple Linear Model model.



Results for Multiple Linear Regression Model

At the end of my analysis, I came up with the following formula:

Results for Multiple Linear Regression Model

At the end of my analysis, I came up with the following formula:

$$
\begin{aligned}
\ln(\mathit{price}) = & 10.2082 + 0.3618 \cdot \text{waterfront} - 0.0160 \cdot \text{bedrooms} \\
& - 0.0153 \cdot \text{bathrooms} + 0.1400 \cdot \text{sqft\_living}^{0.3} + 0.0088 \cdot \text{floors} \\
& + 0.1494 \cdot \text{view}^{0.5} + 0.0105 \cdot \text{grade}^{2} + 0.1187 \cdot \ln\left(\text{sqft\_living15}\right)
\end{aligned}
$$

Check Statistical Hypotheses

This Linear Model satisfies all statistical assumptions of the Linear Regression, namely:

- Linearity.
- Normality.

The model is very close to satisfy Constant Error Variance.

Overall Conclusion:

I conclude that our model almost satisfies statistical assumptions for the regression model.

Model

$$
\begin{aligned}
\ln(\textit{price}) =& 10.2082 + 0.3618 \cdot \text{waterfront} - 0.0160 \cdot \text{bedrooms} \\
& - 0.0153 \cdot \text{bathrooms} + 0.1400 \cdot \text{sqft\_living}^{0.3} + 0.0088 \cdot \text{floors} \\
& + 0.1494 \cdot \text{view}^{0.5} + 0.0105 \cdot \text{grade}^{2} + 0.1187 \cdot \ln\left(\text{sqft\_living15}\right)
\end{aligned}
$$

Explanation of the Model

Before I begin explain the coefficients, I notice that $P > |t|$ for the *floors* variable is 0.155, which makes *floors* insignificant for the analysis.

Intercept

- The model has the sample intercept of 10.2082.

Intercept

- The model has the sample intercept of 10.2082.
- If we assume that all explanatory variables are zeros, this would mean that the price would be $e^{10.2082} \approx 27,124$

0.3618 is the coefficient for *waterfront*

Waterfront is a categorical variable coded as 0 or 1.

0.3618 is the coefficient for *waterfront*

Waterfront is a categorical variable coded as 0 or 1.

To understand the change in *price* in percents, if we switch from the house with no waterfront to the house with waterfront while keeping all other variables the same, will use the following formula:

$$100 \times \left[e^{0.3618} - 1\right] \approx 43.59\%$$

0.3618 is the coefficient for *waterfront*

Waterfront is a categorical variable coded as 0 or 1.

To understand the change in *price* in percents, if we switch from the house with no waterfront to the house with waterfront while keeping all other variables the same, will use the following formula:

$$100 \times \left[e^{0.3618} - 1\right] \approx 43.59\%$$

Switching from the house with no waterfront to the house with waterfront while keeping all other explanatory variables fixed, will increase the price by 43.59%.

0.1400 is the coefficient for *sqft_living*

If we increase the *sqft_living* from $1000ft$ to $1100ft$ while keeping the other variables fixed, we will get the following change in price in percents:

$$100 \times \left[e^{0.1400\left(1100^{0.3}-1000^{0.3}\right)}-1\right] \approx 3.28\%$$

- If the *sqft_living* is $1000ft$ and we increase it to $1100ft$ while keeping the other variables fixed, we will get the change in price of 3.28%.

0.1400 is the coefficient for *sqft_living*

If we increase the *sqft_living* from 1000*ft* to 1100*ft* while keeping the other variables fixed, we will get the following change in price in percents:

$$100 \times \left[e^{0.1400\left(1100^{0.3} - 1000^{0.3}\right)} - 1\right] \approx 3.28\%$$

- If the *sqft_living* is 1000*ft* and we increase it to 1100*ft* while keeping the other variables fixed, we will get the change in price of 3.28%.
- In this particular example 10% change in *sqft_living* starting from 1000*ft* forces 3.28% change in price.

0.1494 is the coefficient for the *view*

The coefficient for the *view* has the same explanation as the *sqft_living*.

0.1494 is the coefficient for the *view*

The coefficient for the *view* has the same explanation as the *sqft_living*.

If we increase the *view* by 1 unit from 2 to 3 while keeping the other variables fixed, we will get the following:

$$100 \times \left[e^{0.1494\left(3^{0.5}-2^{0.5}\right)} - 1\right] \approx 4.86\%$$

In this particular example if the *view* will increase from 2 to 3, the price will increase by 4.86%.

0.0105 is the coefficient for the *grade*

The coefficient for the *grade* has the same explanation as the *sqft_living*.

0.0105 is the coefficient for the *grade*

The coefficient for the *grade* has the same explanation as the *sqft_living*.

If we increase the *grade* by 1 unit from 2 to 3 while keeping the other variables fixed, we will get the following:

$$100 \times \left[e^{0.0105\left(3^2 - 2^2\right)} - 1 \right] \approx 5.39\%$$

In this particular example if the *grade* will increase from 2 to 3 while other variables stay the same, the price will increase by 5.39%.

0.1187 is the coefficient for the *sqft_living15*

The change in price in percent will be:

$$100 \times \left[\left(\frac{\text{price } 2}{\text{price } 1} \right)^{0.1187} - 1 \right]$$

In this particular example, if we increase the *sqft_living15* by 100 units from 1000 to 1100 while keeping the other variables fixed, we will get the following:

$$100 \times \left[e^{0.1187} \times \frac{1100}{1000} - 1 \right] \approx 1.14\%$$

Thus, 10% increase in *sqft_living15* will lead to 1.14% increase in *price*.

R^2

- The model has $R^2 \approx 0.5$.

R^2

- The model has $R^2 \approx 0.5$.
- This means that our model explains about 50% of the variation by using *sqft_living* as independent variable.

ANOVA

Is our model with many explanatory variable better than the model with zero explanatory variables?

- Our p-value for this model is $p = 0.000 < 0.05 = \alpha$.
- We conclude that the Test tells us, that at least one of the coefficients is not 0.
- Since our p-value is 0, there is a 0% probability that the improvements that we are seeing with our independent variables model are due to random chance alone.

Conclusions: Data Modeling

I used the following steps during data modeling:

- Dropped 2376 rows where *waterfront* has no value.
- Dropped 63 rows where *view* is has no value.
- Dropped 3842 rows where *yr_renovated* has no value.
- I converted *sqft_basement* string format into numeric values.
- During modeling I dropped very large and very small values when necessary.

Conclusions: Modeling

- I built two models:
  - Linear Regression Model
  - Multiple Linear Regression Model
- I checked whether the models satisfy statistical assumptions of Linear Regression
- I explained the models.
- Models can be used for interpolation given the data about a particular property.

Conclusions: Ways to Improve the Analysis

- More data wrangling is need to remove *heteroscedasticity* from the Multiple Linear Regression Model.
- Include more explanatory variables.
- Scrape webpages for more data such as school grade, crime rate, etc. for properties.

THE END

THANK YOU!