# Correlation and Regression

CS51: Formal Analyses
Minerva Schools at KGI

Steven H. Yang
Feb 07, 2021



**Word Count**: 1,400

# Correlation and Regression

## Introduction

From 2001 to 2017, there is a 146% increase in the number of Americans who cannot afford their housing (Popken, 2017). Housing is one of the necessities of life and provides stability in our life. Households consider their housings based on the number of family members because the house should have enough room for them.

This paper analyzes correlation and regression between the prices of houses and their living area so that the possible home-buyers in King County, WA can predict and make a better purchase in the future.[1]

Based on the descriptive statistics, correlation coefficient will be calculated and analyzed. After this, this paper discusses the regression analysis with relevant formulas to test the hypotheses.

## Dataset

Kaggle provides house sales in King County, WA data between May 2014 and May 2015. This data set includes the house price and its number of bedrooms, the area of living in square feet, etc. This report analyzes the correlation between the price of the house and its living area.

Two hypotheses will be tested in this study:

- H0: B = 0, There is no linear relationship between the price of a property and the area of its living space.

[1] **#purpose:** Effectively and clearly explains the goal of the research and how this research would be valued.

- HA: $B > 0$, There is a positive relationship between the price of a property and the area of its living space.

Based on the correlation, this report expects a linear regression line to foresee the price of the house in king county. With the slope of the regression line, these hypotheses would be tested.

Multiple variables help the analysis. Since both prices and living areas are nearest hundreds of dollars and the nearest tens of feet, they are discrete quantitative variables. Assume there is a linear association between the prices and the living area; a linear regression line is a helpful tool to predict the house price based on the living area. In this case, the living area is a predictor variable, and the price is a response variable. These variables are critical to analyze. The result of this study will be helpful for the consumers since they can consider how big the house is and its affordability. Therefore, it will help consumers who are looking for a spacious and affordable house. The imported data can be viewed in Appendix A.[2]

**Methods I (Correlation)**

The current data set includes 21,613 houses in King County. There are several houses from Seattle in this data, which may result in outliers. These outliers can be observed in Figure 3. To satisfy the absence of outliers assumption, this paper eliminates these outliers. The definition of outliers is any data points that are three standard deviations away from the mean. Following descriptive statistics and distribution are excluding the outliers.

[2] **#variables:** Clearly and effectively explains the variables and their importance in this research. Explains how the variables will be used in #regression.

| **Table 1: Summary Statistics for Each Variable after the Elimination of Outliers**[3] | | |
|---|---|---|
| | House Prices ($ in millions) | Living Areas (sqft) |
| Count | $n_1 = 21,088$ | $n_2 = 21,088$ |
| Mean | $\bar{x}_1 = 503,113.84$ | $\bar{x}_2 = 2012.41$ |
| Median | 445,000.00 | 1890.00 |
| Standard Deviation | $s_1 = 254316.45$ | $s_2 = 796.14$ |

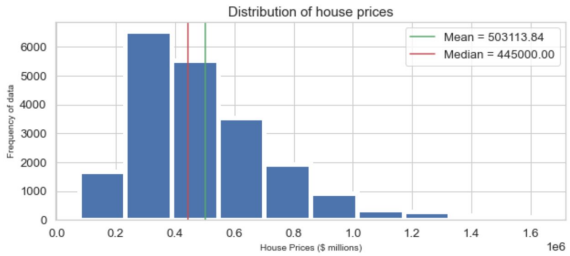


**Figure 1: Histogram for House Prices in King County. The histogram is still right-skewed after the outlier removal.**

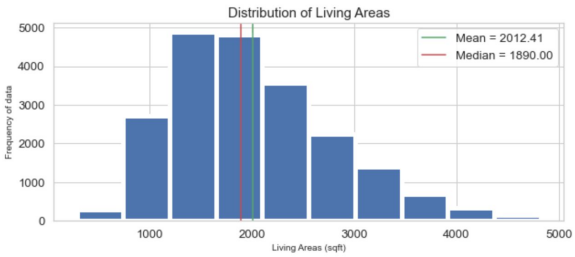


**Figure 2: Histogram for Living Areas of Houses in King County The histogram is still right-skewed after the outlier removal.**

Both histograms have lower median values than their mean values. This means they are both right-skewed. Even after dropping outliers off, some houses have much higher prices than most of the houses. This paper considered defining outliers as any data points that are two standard deviations away from the mean instead of three because it gives a more normally distributed

[3] **#descriptivestats:** Calculated appropriate descriptive statistic only that will be used for #correlation. Clearly and correctly created histograms and interpreted the statistics effectively.

shape for the histograms. It would possibly help the future analyses of the population; however, this paper will not do this because it eliminates more than 5% of the current dataset, which means it can eliminate essential data.[4] Python code for Table 1, Figure 1, and 2 can be found in Appendix B and C.

This paper analyzes the correlation between the area of living space and houses' price for both before/after outlier removal for comparison. As one can find in Figure 3 and Figure 4, the Pearson's r-value has been decreased after the removal. This is an unexpected result; however, this is still a possible situation and helps us understand the correlation between the living area and prices more accurately. Python code for Table 1, Figure 3, and 4 can be found in Appendix D.

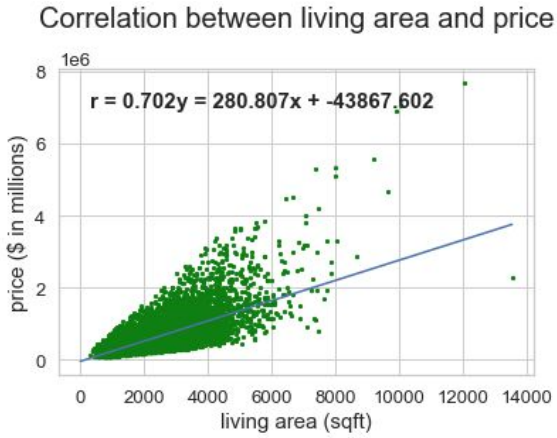


**Figure 3: Correlation before Outlier Removal. There are multiple outliers.**

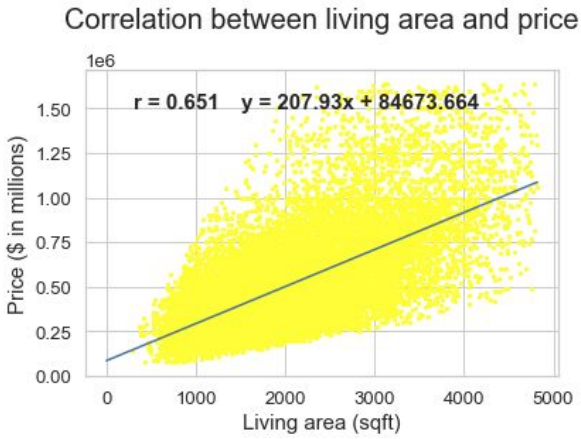


**Figure 4: Correlation after Outlier Removal. R-value has been decreased after outlier-removal but it represents the real-world model better.**

Pearson's r-value can be calculated with the following formula:

[4] **#distributions:** Appropriately identified the distribution type and its importance. Explained how outliers could be removed more based on the distribution. Effectively explained the possible results with normal distribution.

$$r = \frac{1}{n}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

Inside of the sigma, each x and y terms can be replaced with $z_x$ and $z_y$. It means the formula analyzes the distance between the point and mean with their z-scores so that they can evaluate how much they are correlated. Please refer to Table 1 for the variable definition of the formula. $x_i$ and $y_i$ mean the initial value. Calculation reported r = 0.651; there is a moderate positive association between the living area and houses' price.[5]

**Methods II (Regression)**

This paper has five assumptions to apply linear regression: linear, independent, normal residuals, the equal variance of residuals, and random. These will be verified later in this paper.

[5] **#correlation:** Based on the appropriate #descriptivestats, calculated the accurate Pearson's r coefficient. Effectively interpreted how the data are associated.

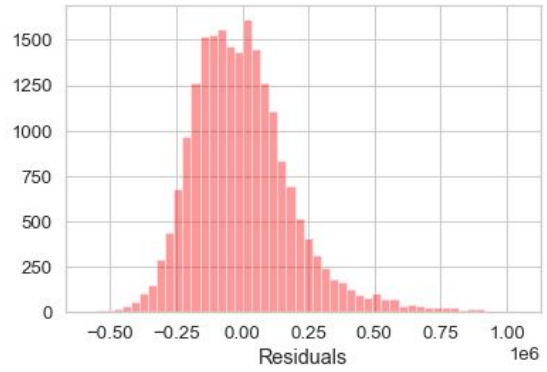


**Figure 5: Normality of Residuals. The distribution is nearly normal; residuals are centered at almost zero.**

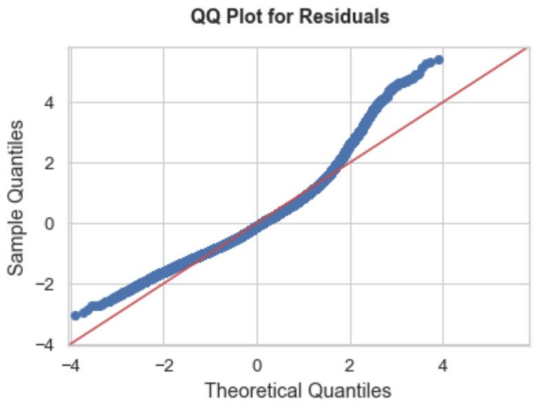


**Figure 6: QQ Plot for Residuals. Most of the points of QQ plot are closely located from the line: nearly equal variance.[6]**

Figure 5 tells that residuals are nearly-normally distributed because the residuals are centered at nearly zero. Figure 6 tells that residuals share nearly equal variance to each other based on the homoscedasticity. From this, the assumptions have been verified. Python code for these figures can be found in Appendix E.

One can expect a linear regression line between the predictor variable and the response variable due to the correlation. R-squared value can be calculated with the following formula:

$$r^2 = 1 - \frac{Unexplained\ Variation\ (SSE)}{Total\ Variation\ (SSTO)}$$

As seen in the formula, r-squared value measures how many points can be explained by the regression model by subtracting unexplained points. The calculation reported the r-squared value as 0.424. It means 42.4% of the data can be explained with the linear regression model: Price ($) = 207.93 * Living Area (sqft) + 84673.664. 207.93 is a slope coefficient of the model and

[6] **#dataviz:** From Figure 1 to Figure 6, clearly stated the x and y axis. Explained the importance of each data visualization in the caption. Drew reader-friendly data visualizations to provide better understandings.

84573.664 is a constant. Therefore, per one square-feet, there is an increase of $207.93 starting with $84673.664.

| **Table 2: OLS Regression Results** | | | |
|---|---|---|---|
| **Dep. Variable:** | Price ($ in millions) | **R-squared:** | 0.424 |
| **Model:** | OLS | **Adj. R-squared:** | 0.424 |
| **Method:** | Least Squares | **Df Model:** | 1 |
| **No. Observations:** | 21088 | **Df Residuals:** | 21086 |

| **Table 3: OLS Regression Results (continued)** | | | | | | |
|---|---|---|---|---|---|---|
| | **Coefficient** | **std err** | **t** | **P>\|t\|** | **[0.025** | **0.975]** |
| **Constant** | 8.467e+04 | 3614.130 | 23.429 | 0.000 | 7.76e+04 | 9.18e+04 |
| **Living Area (sqft)** | 207.9297 | 1.670 | 124.510 | 0.000 | 204.656 | 211.203 |
| **Skew:** | | 1.010 | **Prob(JB):** | | | 0.00 |

Based on Table 3, 95% of the constant are between 77,600 and 91,800, and 95% of the Living Area (sqft)'s coefficient is between 204.656 and 211.203. Therefore, one can predict the house's price with the area of living space in 95% of chance by using the values between confidence intervals with the linear regression model. Python code for the calculation table can be found in Appendix F.

P-value means the probability that one can observe that the t-score is equal to or greater than the calculated t-score when the null hypothesis is true. If the p-value is lower than the significance level, one can reject the null hypothesis. Set alpha as 0.05, p-value is almost zero which means this paper rejects the null hypothesis.[7]

Since the slope's confidence interval [204.656, 211.203] means the slope would be a value between the interval in 95% of chance, it is consistent with the hypothesis testing above.[8] Therefore, there is a positive relationship between the house prices and their living areas.[9]

**Results and Conclusions**

It promises that this research can reject the null hypothesis as this paper rejected with slope-analysis. This research provides statistically significant evidence to support this conclusion.

This research gives one linear regression model for predicting the house price in King County based on the area of living space. Assume consumers don't put any outlier variable as a predictor variable to the regression model, one can predict their house price correctly with a nearly 42.4%

[7] **#significance:** Set alpha equal to 0.05 to have 95% confidence. Compared alpha with the p-value to process the hypothesis testing.
[8] **#confidenceintervals:** Calculated the confidence interval of the slope and compared with the hypothesis testing result. Based on the confidence intervals, hypothesis testing gets stronger.
[9] **#regression:** Based on the scatter plots, accurately created the regression model. Verified all the assumptions in this research with residuals and their QQ plot. Interpreted the meaning of the regression model accurately.

chance. However, they can improve the accuracy by changing the coefficient and constant of the model based on the confidence interval defined above.[10]

This paper provides a regression model for the future consumers who plan to buy a home in King County, WA. However, this research has a few limitations that should be considered:

1. The dataset includes the houses in Seattle, WA where generally houses are more expensive than any other areas in the county. Even though this paper tried to remove outliers, there are still properties in Seattle that remain in the updated dataset.
2. This research is King County centered; therefore, cannot be used as analyzing the national level even if the research was motivated by the national housing issues.

This research expects the better regression model if this research was conducted separately between Seattle and the King County excludese Seattle. Since urban, suburban and rural areas have different types of housing and values, this research can be done under these three categories as well. In this case, consumers can choose the regression model based on the specific area they are interested in and may get a better prediction.

**Reflection**

Throughout the unit, I learned that the world view can be represented with statistical analyses. From these, one can interpret and predict their idea to move on decision-making.

[10] **#induction:** Based on the regression model and the assumptions of consumers using the model, explained the premises of the interpretation. From premises, successfully explained one's chance of getting prediction correctly with inductive reasoning.

Correlation/Regression are the tools for understanding what the data says while Significance/Confidence Intervals are knowledge diagnosing how we can use Correlation/Regression practically to our real-life situations. One can interpret the data and can make a decision based on how confidence they feel from the statistics and can use this for their cost-benefit analysis and risk considerations.

## References[11]

Diez, D., Cetinkaya-Rundel, M., & Barr, C. (2015). *OpenIntro statistics - Third edition*. Open Textbook Library.

Popken, B. (2017, July 6). *Americans Who Can't Afford Their Homes Up 146 Percent*. NBC News. https://www.nbcnews.com/business/real-estate/americans-who-can-t-afford-their-homes-146-percent-n774106

[11] **#professionalism:** Used Grammarly as to make sure there are no grammar issues. Proof-read by three friends to detect any errors. Followed the APA guideline for the reference. Formatted with constant standards.

# Appendix

## *Appendix A*

```
#import all the necessary libraries

%matplotlib inline
import pandas as pd
import numpy as np
from scipy import stats
#from scipy.stats import linregress
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(color_codes=True, font_scale = 1.2)
sns.set_style("whitegrid")
plt.rcParams.update({'font.size': 14})
import os #install statsmodels
os.system('pip install statsmodels')
import statsmodels.api as statsmodels # useful stats package with regression functions
import statsmodels.formula.api as smf


#import the data set

df = pd.read_csv('kc_house_data.csv')
print(df)
data_frame = pd.DataFrame(['Living Area (sqft)','Price ($ in millions)'])
price = df["Price ($ in millions)"]
living_area = df["Living Area (sqft)"]

z_scores = stats.zscore(df) #calculate z-scores of df'

abs_z_scores = np.abs(z_scores)
filtered_entries = (abs_z_scores < 3).all(axis=1)
new_df = df[filtered_entries] #eliminate the outliers that are more than three SDs away

print(new_df)
new_df.head()
new_price = new_df["Price ($ in millions)"]
new_living_area = new_df["Living Area (sqft)"]
```

*Appendix B*

```
new_price_mean = round(new_price.mean(),2) #round up to two digits
new_price_median = new_price.median()
sd_new_price = round(np.std(new_price),2) #round up to two digits
new_living_area_mean = round(new_living_area.mean(),2)
new_living_area_median = new_living_area.median()
sd_new_iving_area = round(np.std(new_living_area),2)

def mode(lst): #since I have multiple modes, I created another function for mode
    L1=[] #count the occurrence of each number and append or add findings to the list

    i = 0 #count the numbers and put them into L1
    while i < len(lst) :
        L1.append(lst.count(lst[i]))
        i += 1

    # the occurrences for each number in sorted lst
    # create a custom dictionary d1 for k : V
    # k = value, v = occurence

    d1 = dict(zip(lst, L1))
    d2=[k for (k,v) in d1.items() if v == max(L1)] # the k values with the highest v values.
    return d2

print("Your # of data for price is:",len(new_price))
print("Mean for price is:", price_mean)
print("Median for price is:", price_median)
print("Mode for price is:", sorted(mode(price.to_list())))
print("Standard Deviation for price is:", sd_price)
print("Your range for price is:", max(price) - min(price))
print("######")
print("Your # of data for living_area is:",len(new_living_area))
print("Mean for living_area is:", living_area_mean)
print("Median for living_area is:", living_area_median)
print("Mode for living_area is:", sorted(mode(living_area.to_list())))
print("Standard Deviation for living_area is:", sd_living_area)
print("Your range for living_area is:", max(living_area) - min(living_area))
```

*Appendix C*

```
plt.hist(new_price, alpha=1, linewidth=5)
plt.title("Distribution of house prices", fontsize = 15)
plt.xlabel("House Prices ($ millions)", fontsize = 10)
plt.ylabel("Frequency of data", fontsize = 10)
plt.gcf().set_figwidth(10)

plt.axvline(new_price_mean, color='g', label = "Mean = 503113.84") #label legend for mean
plt.axvline(new_price_median, color='r', label = "Median = 445000.00")#label legend for median
plt.legend()
ax=plt.gca()
ax.set_facecolor('w')
plt.show()
```

```
plt.hist(new_living_area, alpha=1, linewidth=5)
plt.title("Distribution of Living Areas", fontsize = 15)
plt.xlabel("Living Areas (sqft)", fontsize = 10)
plt.ylabel("Frequency of data", fontsize = 10)
plt.gcf().set_figwidth(10)

plt.axvline(new_living_area_mean, color='g', label = "Mean = 2012.41") #label legend for mean
plt.axvline(new_living_area_median, color='r', label = "Median = 1890.00")#label legend for median
plt.legend()
ax=plt.gca()
ax.set_facecolor('w')
plt.show()
```

*Appendix D*

```
def scatter_plot1(x, y, title, color): #define the scatter plot function
    plt.figure()
    plt.scatter(x, y, s=5, c=color)
    plt.title(title, fontsize=20,y=1.2, pad=-14)
    plt.xlabel('living area (sqft)', fontsize=15)
    plt.ylabel('price ($ in millions)', fontsize=15)
    xmax = max(x) #get the maximum value of x
    ymax = max(y) #get the maximum value of y

    #calculation
    slope, intercept, r_value, p_value, std_err = linregress(x, y)
    plt.plot([0, xmax], [intercept, slope * xmax + intercept]) #plot the equation

    # adding legend
    equation = 'y = ' + str(round(slope,3)) + 'x' + ' + ' + str(round(intercept,3))
    rvalue = 'r = ' + str(round(r_value,3))
    plt.text(3000, 7000000, equation,fontsize=15,fontweight='bold')
    plt.text(300, 7000000, rvalue,fontsize=15,fontweight='bold')
    #This python code is adpated from CS51 Session 1.2 and revised by Steven Yang.
def scatter_plot2(x, y, title, color):
    plt.figure()
    plt.scatter(x, y, s=5, c=color)
    plt.title(title, fontsize=20,y=1.2, pad=-14)
    plt.xlabel('Living area (sqft)', fontsize=15)
    plt.ylabel('Price ($ in millions)', fontsize=15)
    xmax = max(x)
    ymax = max(y)

    #calculation
    slope, intercept, r_value, p_value, std_err = linregress(x, y)
    plt.plot([0, xmax], [intercept, slope * xmax + intercept])


    # adding legend
    equation = 'y = ' + str(round(slope,3)) + 'x' + ' + ' + str(round(intercept,3))
    rvalue = 'r = ' + str(round(r_value,3))
    plt.text(1500, 1500000, equation,fontsize=15,fontweight='bold')
    plt.text(300, 1500000, rvalue,fontsize=15,fontweight='bold')
    # second function is needed to fix the location of rvalue and equation
```

*Appendix E*

```
def regression_model(column_x, column_y):

    X = statsmodels.add_constant(new_df[column_x])
    Y = new_df[column_y]
    regressionmodel = statsmodels.OLS(Y,X).fit() #"ordinary lease squares"

    #Getting and calculating relevant values and round them up to 3 decimal points.
    Rsquared = round(regressionmodel.rsquared,3)
    slope = round(regressionmodel.params[1],3)
    intercept = round(regressionmodel.params[0],3)

    #plotting them into #dataviz
    fig, (ax1, ax2) = plt.subplots(ncols=2, sharex=True, figsize=(12,4))
    sns.regplot(x=column_x, y=column_y, data=new_df, marker="x", ax=ax1, color = 'g', scatter_kws={"s": 0.1}) # scatter
    sns.residplot(x=column_x, y=column_y, data=new_df, ax=ax2, scatter_kws={"s": 1}) # residual
    ax2.set(ylabel='Residuals')
    ax2.set_ylim(min(regressionmodel.resid)-1,max(regressionmodel.resid)+1)
    plt.figure()
    sns.distplot(regressionmodel.resid, kde=False, axlabel='Residuals', color='red') # histogram
    qqplot = statsmodels.qqplot(regressionmodel.resid,fit=True,line='45') #QQ plot created
    qqplot.suptitle("QQ Plot for Residuals",fontweight='bold',fontsize=14)

    #print the calculations and regression model
    print("R-squared = ",Rsquared)
    print("Regression equation: "+column_y+" = ",slope,"* "+column_x+" + ",intercept)
    #This python code is adpated from CS51 Session 1.2 and session 2.2. Revised by Steven Yang'24.
```

*Appendix F*

```
global regressionmodel
regressionmodel = statsmodels.OLS(new_df['Price ($ in millions)'],statsmodels.add_constant(new_df['Living Area (sqft)']
regressionmodel.summary() #print the summary of OLS Regression Results
#This python code is adpated from CS51 Session 2.2 and revised by Steven Yang.
```