AMPBA Batch 15 – 2021 Winters

ISB

# Statistical Analysis - II Group Assignment

**Group Submission by**

- Mohua Sinha (12020015)
- Ila Barshilia (12020022)
- Nishant Jalasutram (12020051)
- Deep Kamal Singh (12020053)

Subject Professor – Dr. Raghuram Bommaraju

## Table of Contents

April 13, 2021

# Questions

## Part A

A real estate agent is trying to understand the nature of housing stock and home prices in and around a medium sized town in upstate New York. She has collected data from a random sample of 1047 homes sold in the last 12 months. Data was collected on the following variables and is available in the houseprices.csv file.

- *Price* – the sale price of the house in *$*
- *Living Area* – in *Sq. ft*.
- *Bathrooms* – number of bathrooms in the house (powder rooms with no tub or shower area are considered 0.5 baths)
- *Bedrooms* – the number of bedrooms
- *Lot Size* – size of the property on which the house sits (in acres).
- *Age* – of the house in years
- *Fireplace* – whether or not the house has a fireplace (Yes = 1, No = 0)

Your task in this assignment is to analyse this dataset to gain some understanding of this particular real estate market – the values of homes, their characteristics in terms of size and other features, and relationships between these. This understanding will prove immensely helpful to the real estate agent in advising her clients. Since all of the homes are from the same geographical area, location (which usually has a huge bearing on home values) is not a major concern here.

Most of the analysis will be done in response to the specific questions posed on the homework assignments. But feel free to explore and play around with the data set to enhance your own understanding of how to make sense of data.

1. a) Your friend claims that the average house price in this area is above $150K. Do you agree? Briefly explain what the p-values in these cases mean?(1mark)

   b) He also claims that the average living area is more than 1800 Sq. Ft. Do you agree with this? (Use a 5% significance level for both.). Briefly explain what the p-values in these cases mean? (1mark)

2. Are the home prices higher for houses with fireplaces as compared to those without?

   a) Create side-by-side box plots of the house prices of the two groups and comment them. (2marks)

   b) Formulate an appropriate hypothesis and test it in order to check the above claim. Assume that the population standard deviations of house prices in the two groups are equal. (1mark)

3. Any house aged more than 30 years is considered an "old" house. Your friend claims that old houses have larger lot sizes than new houses. Do you agree? Explain. Use a significance level of 5% for your test. Historical data suggests that old houses include some very large and some very small lot sizes but new houses are more homogeneous in their lot sizes. (2marks)

4. Based on the evidence available here, would you be willing to claim that fireplaces have become more fashionable? For simplicity, it is OK to compare only "new" houses and "old" houses. Use a significance level of 5% for your test. Use a significance level of 5% for your test. (1mark)

5. Suppose that houses with 1-2 bedrooms are considered to be "Small Houses", those with 3-4 are "Medium Houses" and 5-6 as "Big Houses". Can we conclude that the prices of Small, Medium and Big houses are not the same, at 1% level of significance? (2marks)

## PART B

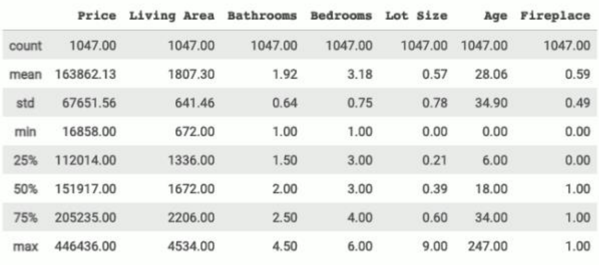
The data for this problem is available in Apple Advertising Data.xlsx, which provides Apple's revenues and advertising expenses.

1. Run a regression of revenues on advertising. Is advertising having a significant impact on revenues? Interpret the intercept and slope. (2 points)
2. Construct a residual plot. Do the residuals appear random? Do they satisfy the first three assumptions? (2 points)
3. Construct a Q-Q plot. Do the residuals appear to be normally distributed? (2 points)
4. If advertising expenses were to be 1500 in 2016, what would your prediction interval for revenues? (2 points)
5. Transform revenues and advertising into a log scale. Repeat the above steps 1-4. (5 points).
6. Would you recommend a linear model or log-log model? Please explain. (2 points)

# Answers

## PART A

### Data Analysis

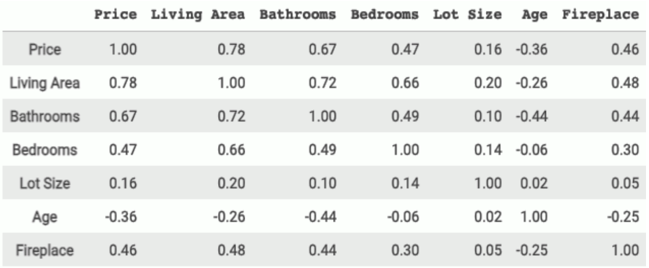


| | Price | Living Area | Bathrooms | Bedrooms | Lot Size | Age | Fireplace |
|---|---|---|---|---|---|---|---|
| count | 1047.00 | 1047.00 | 1047.00 | 1047.00 | 1047.00 | 1047.00 | 1047.00 |
| mean | 163862.13 | 1807.30 | 1.92 | 3.18 | 0.57 | 28.06 | 0.59 |
| std | 67651.56 | 641.46 | 0.64 | 0.75 | 0.78 | 34.90 | 0.49 |
| min | 16858.00 | 672.00 | 1.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| 25% | 112014.00 | 1336.00 | 1.50 | 3.00 | 0.21 | 6.00 | 0.00 |
| 50% | 151917.00 | 1672.00 | 2.00 | 3.00 | 0.39 | 18.00 | 1.00 |
| 75% | 205235.00 | 2206.00 | 2.50 | 4.00 | 0.60 | 34.00 | 1.00 |
| max | 446436.00 | 4534.00 | 4.50 | 6.00 | 9.00 | 247.00 | 1.00 |

**Table 1.** Five-point summary of Dataset

**Conclusion we draw from Five-point summary of each data point**

- The data set contains 1047 records.
- None of records have null value for any data point.
- Average house price as per sample is ~$164K
- Average Living area is 1807sq.ft in the sample
- All houses in sample have bathroom and bedroom, on an average there are roughly 2 bathrooms and 3 bedrooms
- Interesting point to note that sample standard deviation is more than average for lot size



| | Price | Living Area | Bathrooms | Bedrooms | Lot Size | Age | Fireplace |
|---|---|---|---|---|---|---|---|
| Price | 1.00 | 0.78 | 0.67 | 0.47 | 0.16 | -0.36 | 0.46 |
| Living Area | 0.78 | 1.00 | 0.72 | 0.66 | 0.20 | -0.26 | 0.48 |
| Bathrooms | 0.67 | 0.72 | 1.00 | 0.49 | 0.10 | -0.44 | 0.44 |
| Bedrooms | 0.47 | 0.66 | 0.49 | 1.00 | 0.14 | -0.06 | 0.30 |
| Lot Size | 0.16 | 0.20 | 0.10 | 0.14 | 1.00 | 0.02 | 0.05 |
| Age | -0.36 | -0.26 | -0.44 | -0.06 | 0.02 | 1.00 | -0.25 |
| Fireplace | 0.46 | 0.48 | 0.44 | 0.30 | 0.05 | -0.25 | 1.00 |

**Table 2.** Correlation matrix of Dataset

**Correlation Analysis**

**As expected:**

- Living area and price are strongly positively correlated followed by bathrooms, bedrooms and fireplace. Lot size has the weakest correlation with the price among all variables.
- Age has negative correlation. The prices decrease as the houses age more.
- Larger the living area, more the no. of bathrooms and bedrooms.

**Counter Intuitive:**

- Age has no relation with Lot size, no. of Bedrooms.
- Fireplace is negatively correlated with age. As age becomes more, lesser the houses with fireplace which could indicate that fireplace may be a fashion trend in newer houses. to understand this more, this data needs further analysis.

## Solution to PART A

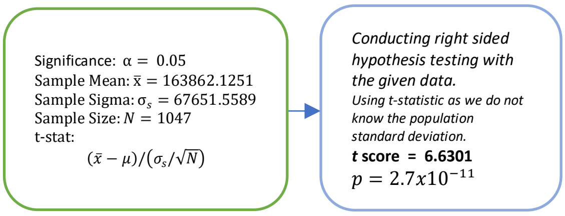
### *1.a*

Based on “friend’s claim” and observed sample, we build our hypothesis as:

**Null Hypothesis**: Average housing price is less than or equal to $150K, and **Alternate Hypothesis**: Average housing price is more than $150K

$$\therefore \quad H0 : \mu <= \$150{,}000$$
$$H1 : \mu > \$150{,}000$$

We formulate:



Significance: $\alpha = 0.05$
Sample Mean: $\bar{x} = 163862.1251$
Sample Sigma: $\sigma_s = 67651.5589$
Sample Size: $N = 1047$
t-stat:
$(\bar{x} - \mu)/(\sigma_s/\sqrt{N})$

*Conducting right sided hypothesis testing with the given data.*
*Using t-statistic as we do not know the population standard deviation.*
***t* score = 6.6301**
$p = 2.7x10^{-11}$

- Corresponding to *t* score of 6.6301 the p-value is 0.000000000027 which is less than significance of 0.05 **thus we are Rejecting null hypothesis.**
- The *p*-value is extremely low which indicates that by rejecting the null hypothesis we will be committing a very insignificant Type-1 error
- This indicates a very strong support to my friend's claim that the average housing price is greater than $150,000

## 1.b

Based on “friend’s claim” and observed sample, we build our hypothesis as:

**Null Hypothesis**: Average living area is less than or equal to 1800sq.ft, and **Alternate Hypothesis**: Average living area is more than 1800sq.ft

$$\therefore \quad H0: \mu \leq 1800 sq.ft$$
$$H1: \mu > 1800 sq.ft$$

We formulate:

Significance: $\alpha = 0.05$
Sample Mean: $\bar{x}$ = 1807.30
Sample Sigma: $\sigma_s = 641.4609$
Sample Size: $N = 1047$
*t*-stat:
$(\bar{x} - \mu)/\sigma_s/\sqrt{N}$

*Conducting right sided hypothesis testing with the given data.*
*Using t-statistic as we do not know the population standard deviation.*
***t* score = 0.3684**
***p* = 0.3563**

- Corresponding to *t* score of 0.3684 the p-value is 0.3563 which is more than significance of 0.05 **thus we fail to reject the null hypothesis.**
- The p-value that we received from the sample indicates that if we reject the null hypothesis the probability of committing a Type-I error will be 35.63% which is very significant and beyond accepted significance level of 5%
- This indicates that there is not enough evidence to the friend's claim that the average living area is greater than 1800*sq.ft.*

## 2.a

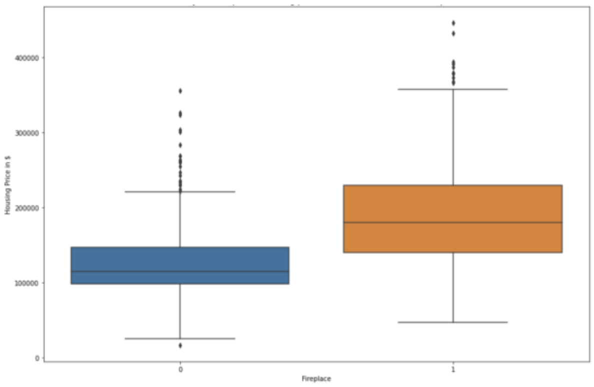


**Figure 1**. Side by Side boxplots of house prices for houses with and without Fireplaces

**As per Figure 1, we conclude:**

1. From the boxplots, we can clearly observe that the data for both the cases is right skewed and there is presence of outliers in both cases
2. Additionally, the median price value of the houses with fireplace is higher than that of houses without fireplace
3. Almost 60-70 percentile of the price values for houses without fireplace fall within the 25th percentile of the values with fireplace

## 2.b

We build our hypothesis (Two sample independent test), It is stated to assume that population standard deviations of house prices in the two groups are equal.

*Let*:

- $\mu_1$ represent the mean housing price of houses with fireplace
- $\mu_2$ represent the mean housing price of houses without fireplace
- $\mu_{D0}$ represent the claimed mean housing price difference between houses with & without fireplace. This is 0 in this case.
- $\alpha$ = 0.05 Assuming the significance level of 0.05

**Hypothesis**

$$\therefore \quad H0: \mu_1 - \mu_2 \leq 0$$
$$H1: \mu_1 - \mu_2 > 0$$

The test statistic for the paired-observation t test is

$$df = N_1 + N_2 - 2 = 1045$$

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \qquad t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{S_P^2(1/n_1 + 1/n_2)}}$$

***t* = 16.7581**

$$p\ value\ of\ t\ score = \mathbf{0.00}$$

- Corresponding to *t* score of 16.7581 the p-value is 0.00 which is very less than the significance of 0.05 thus we are **Rejecting null hypothesis. Claimed mean pricing difference is greater than zero.**
- The p-value is so significantly low that it indicates by rejecting the null hypothesis, we will be committing a very insignificant Type-1 error
- This indicates a very strong support to the claim that the home prices are higher for houses with fireplaces as compared to those without

## 3

We formulate our hypothesis for Two sample independent test with varied population standard deviations, It is stated that the lot sizes differ a lot in old houses and homogeneous in new houses.

This clearly indicates that the population standard deviations for lot size for the old & new houses are not same.

*Let*:

$\boldsymbol{\mu_{new}}$ represent the mean lot size of new houses

$\boldsymbol{\mu_{old}}$ represent the mean lot size of old houses

$\boldsymbol{\mu_{D0}}$ represent the claimed mean lot size difference between old & new houses. This is 0 in this case.

$\alpha$ = 0.05 Assuming the significance level of 0.05

**Hypothesis**

$$\therefore \qquad H0 : \mu_{new} - \mu_{old} \le 0$$
$$H1 : \mu_{new} - \mu_{old} > 0$$

The test statistic for the paired-observation t test is given by

$$\text{df} = \left\lfloor \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{(S_1^2/n_1)^2/(n_1-1) + (S_2^2/n_2)^2/(n_2-1)} \right\rfloor$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

Substituting values we get:

$t = 0.6080$

$p$ value for the *t score*

$= \mathbf{0.271712007448}$

- Corresponding to *t* score of 0.6080 the p-value is 0.2717 which is considerably greater than the significance of 0.05 ,thus we **Fail to reject null hypothesis**. Claimed mean lot size difference may not be greater than zero.
- **The p-value is so high which indicates that by rejecting the null hypothesis, we will be committing a very high Type-1 error**
- This indicates that the average lot size of old houses may not be larger than the new houses

## 4

We formulate our hypothesis for (Two sample independent test for testing difference between two population proportions):
the test needs to be done to check if there are higher proportion of fireplaces in New houses than old houses.

**That is difference of fireplace proportions of houses between Old & New > 0 or not.**

*Let*:

$\boldsymbol{p_{new}}$ represent the sample proportion of fireplaces in new houses

$\boldsymbol{p_{old}}$ represent the sample proportion of fireplaces in old houses

$\boldsymbol{p_{D0}}$ represent the claimed mean lot size difference between old & new houses. This is 0 in this case.

$\boldsymbol{\alpha} = 0.05$ Assuming the significance level of 0.05

**Hypothesis**

$$\therefore \qquad H0 : p_{new} - p_{old} \le 0$$
$$H1 : p_{new} - p_{old} > 0$$

The test statistic for the difference between two population proportions where the null hypothesis difference is zero is:

$$z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Substituting values we get:

$z = 5.631710054280531$

$pvalueofzscore = \mathbf{8.922x10^{-9}}$

- Corresponding to *t* score of 5.6317 the p-value is $\mathbf{8.922x10^{-9}}$ which is considerably lower than the significance of 0.05 ,thus **Rejecting null hypothesis. Claimed mean lot size difference is greater than zero.**
- The p-value is so significantly low which indicates that by rejecting the null hypothesis, we will be committing a very insignificant Type-1 error
- This indicates a very strong support to the claim that the fireplaces have become more fashionable as the proportion of New houses with fireplaces are more than in Old houses.

## 5

For this analysis we set the hypothesis based on Analysis of variance test or ANOVA Test:
We need to check whether

**The average housing prices across Small, Medium and Big houses are same or not.**

*Let*:

$\boldsymbol{\mu_{small}}$ : represent the average housing price for Small houses.

$\boldsymbol{\mu_{medium}}$: represent the average housing price for Medium houses.

$\boldsymbol{\mu_{big}}$ : represent the average housing price for Big houses.

$\boldsymbol{\alpha} = 0.01$ Assuming the significance level of 0.01 (99%)

**Hypothesis**

$$\therefore \qquad H0 : \mu_{small} = \mu_{medium} = \mu_{big}$$
$$H1 : Not\ all\ three\ \ \mu_i\ are\ equal$$

The test statistic for analysis of Variance is

$$F_{(r-1,\, n-r)} = \frac{\text{MSTR}}{\text{MSE}}, \qquad \text{MSE} = \frac{\text{SSE}}{n - r}, \text{ and}$$

$$\text{SST} = \text{SSTR} + \text{SSE} \Rightarrow$$

$$\text{MSTR} = \frac{\text{SSTR}}{r - 1}$$

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{\bar{x}})^2 = \sum_{i=1}^{r} n_i(\bar{x}_i - \bar{\bar{x}})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

Substituting values we get:

$F = 58.7103$

$\boldsymbol{p}\ value\ of\ F\ score = 0.00$

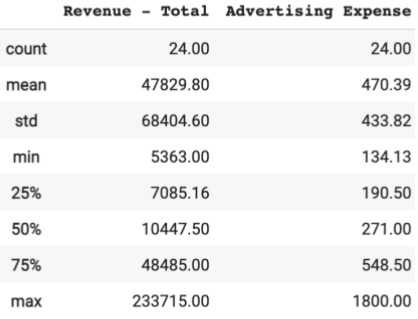
- Corresponding to F score of 58.7103 the p-value is 0.00, thus **Rejecting null hypothesis. Average housing prices across small, medium and big houses are not same.**
- The p-value is so significantly low which indicates that by rejecting the null hypothesis, we will be committing a very insignificant Type-1 error
- This indicates a very strong support that the average housing prices across small, medium and big houses are not same.

# PART B

## Data analysis

1. There are total 24 records – and 2 numerical data point viz. Revenue, and Adv. Exp
2. From 1992 to 2015 – we have total revenue vs advertising expense
3. We see the trend that as time progresses, advertising costs increase and same trend is followed by revenue.
4. Inter quartile values of Revenue and Ad expenses indicate increase in both values YoY

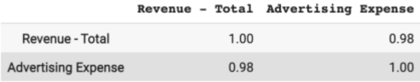
Table 3. Five point summary of data points



| | Revenue - Total | Advertising Expense |
|---|---|---|
| count | 24.00 | 24.00 |
| mean | 47829.80 | 470.39 |
| std | 68404.60 | 433.82 |
| min | 5363.00 | 134.13 |
| 25% | 7085.16 | 190.50 |
| 50% | 10447.50 | 271.00 |
| 75% | 48485.00 | 548.50 |
| max | 233715.00 | 1800.00 |

Table 4 on right suggests that there is strong positive relation between revenue and advertising expenses – as the revenue increases, Ad expenses also increase, this is discussed in more detail in answer to questions

Table 4. Correlation matrix



| | Revenue - Total | Advertising Expense |
|---|---|---|
| Revenue - Total | 1.00 | 0.98 |
| Advertising Expense | 0.98 | 1.00 |

## Solution to Part B

### *1*

- On right we generated plot of regression of revenues on advertising.

*Impact of Advertising on Revenues: Interpreting the intercept and slope:*

- We observe a clear increasing trend from the graph. As advertising spend increases, revenues do too.
- Slope of the equation tells us that for each additional dollar spent on advertising fetches a revenue of $154.39.
- The intercept is negative which mathematically means when advertising spend is 0, the revenue will be -24793.51. However data suggests that all revenues are positive and also practically revenues cannot be a function of advertising alone and fall negative in absence of advertising.
- Hence, practically negative intercept doesn't make sense so we choose to ignore this.

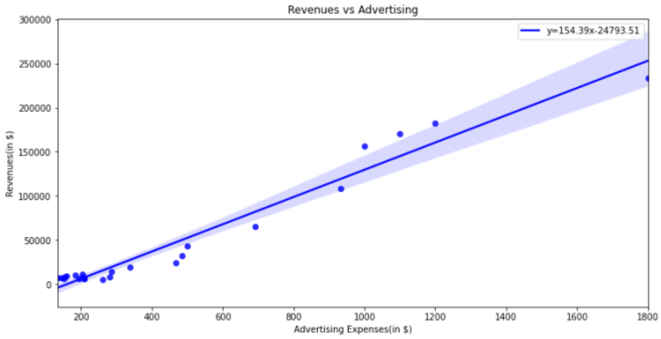


**Figure 2**. Regression plot - Advertising expense vs Revenue

- Slope Equation ⇒ $y = 154.39x - 24793.51$
- Adjusted R Square

### *2*

**Analysis of correlation plot:**

The residuals are not random and independent; they show curvature.

The situation can be corrected by adding the variable X2 to the model. This also entails the techniques of multiple regression analysis

**Assumption 1**: Linearity

From the above Revenues vs Advertising plot, it seems that the relationship of the independent variable with Revenues is not linear

**Assumption 2:** Homoscedasticity

Homoscedasticity means that the residuals have equal or almost equal variance across the regression line.

This quadratic shaped scatter plot obtained from the scatter plot confirms heteroscedasticity is present. (ref probability plot on right), The residuals may show a pattern as predictor variable x is advertising spend each year and hence on x-axis time is increasing. The errors may be correlated as this data is collected over time.

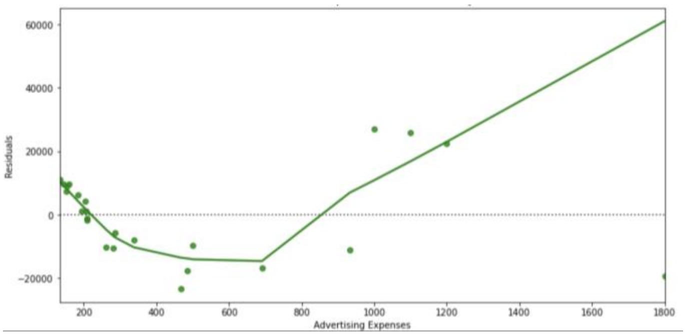


**Figure 3**. Advertising expense vs Residuals plot for Homoscedasticity check

**Assumption 3:** Correlation  Corr[$\varepsilon_i$, $\varepsilon_j$] = 0 for all i ≠ j
correlation between consecutive errors or errors separated by some other number of periods

## 3

**Analysis of Q-Q Plot:**

- The theoretical and actual quantiles of residuals form the Q-Q plot
- In the Q-Q plot, the dots of our plot falls on a 45° line, therefore our data is nearly normally distributed (assuming we are using the normal distribution's theoretical quantiles)

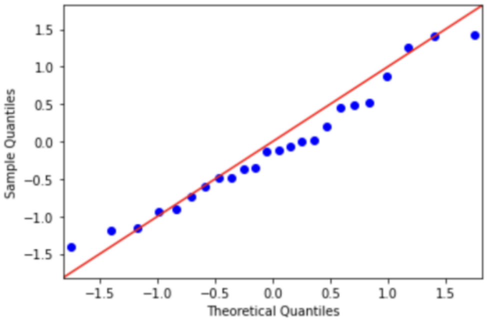


**Figure 4.** Q-Q Plot – Theoretical vs Sample quantities

## 4

| | |
|---|---|
| | |

WORK IN PROGRESS

```
# Prediction Interval  --------------- Reference:
https://online.stat.psu.edu/stat501/lesson/3/3.3
x=data["Advertising Expense"]
x_mean=data["Advertising Expense"].mean()
y=data["Revenue - Total"]
n=data["Advertising Expense"].count()

#calculate standard error of the mean
s=st.sem(y)
SSx=sum((x-x_mean)**2)
tαby2 =2.069
α = 0.025 #95% prediction interval
df=n-1
x1=1500
β0=intercept
β1=slope
ŷ=round((β0+β1*x1),2)
u_interval=ŷ+(tαby2*s*math.sqrt(1+(1/n)+((x1-x_mean)**2)/SSx))
l_interval=ŷ-(tαby2*s*math.sqrt(1+(1/n)+((x1-x_mean)**2)/SSx))
```

For advertising expenses equal to 1500 in 2016,the prediction interval for revenues: $ 173355.84 - $ 240224.84

## 5

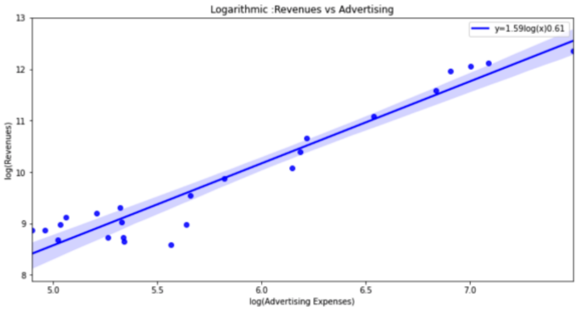
**Part 1**

After transformation to Log-Log scale, we obtain

```
Regression of log revenues on advertising

Model slope:      1.59
Model intercept: 0.61
```

Slope Equation: $\Rightarrow y = 1.59log(x) + 0.61$
Adjusted R Square =

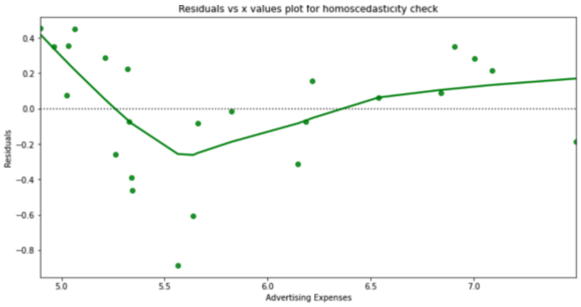


**Part 2**

As per the plot on right, we find that the residuals are not random and independent, they show curvature.

**Assumption 1:** Linearity
From the above Revenues vs Advertising plot, it seems that the relationship of the independent variable with Revenues is linear.

**Assumption 2:** Homoscedasticity
Check for Homoscedasticity. Homoscedasticity means that the residuals have equal or almost equal variance across the regression line.
This above quadratic shaped scatter plot obtained from the scatter plot confirms heteroscedasticity is present.

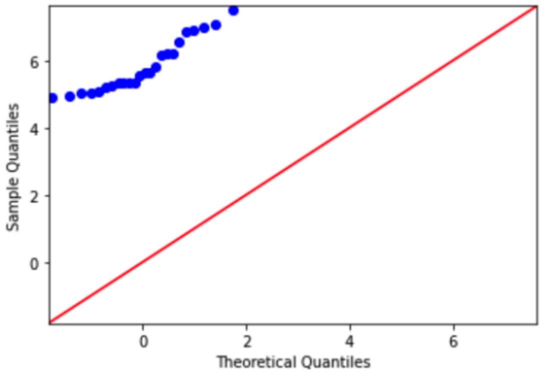


**Assumption 3:** Normality
it is also evident that the residuals are not normal as they align themselves along the curvature in the plot

**Part 3**

When the quantiles of two variables are plotted against each other, then the plot obtained is known as quantile – quantile plot or qqplot.
This plot provides a summary of whether the distributions of two variables are similar or not with respect to the locations.

With logarithmic transformation , still the dots of falls on a line parallel to 45° line, therefore our data is normally distributed.



**Part 4**

## 6

We go ahead to use log-log model generally, when the assumptions for linear model goes wrong; the regression function is not linear and the error terms are not normal and have unequal variances.

1. **Transforming the y values corrects problems with the error terms.**
2. **Transforming the x values primarily corrects the non-linearity.**

When we are not able to establish the assumptions in linear model, we can do a log transform to both the response y and the predictor x.