Project#2: Housing Prices - Inferential Statistics
Majdal Hindi, 300991890

Introduction

In the first project we saw that the area in feet is the **predictor variable (X)** and the housing price in $ is **the response variable (Y)**. We have shown some descriptive statistics and simple analysis of the relationship between the two variables, by using the linear regression model. In this project, we will continue with the same dataset in order to apply some inferential statistics, to use the method of moments to estimate the mean and the variance, to find the confidence intervals and to test the null and the alternative hypothesis.

Inferential Statistics Analysis

- ### Estimation methods

Assuming that Y is normally distributed. By using **Method of Moments:**

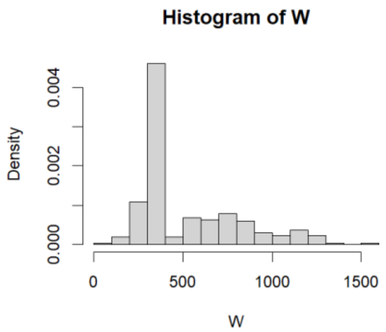
$$E(X_i) = \mu \qquad E(X_i^2) = \sigma^2 + \mu^2$$

mue.hat = 281,171.9$ **sigma2.hat** = (88,952.07$)^2

The minimum value of the Area (**m**) = 410.71 feet. Assuming that W = X – m has Gamma(α,λ) distribution. By using **Method of Moments:**

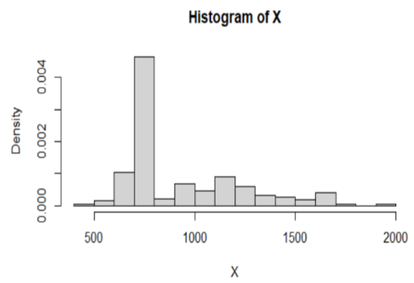
W - statistics:

μ = 525.5113 ft
σ2 = 81165.08 ft^2
Scale = 1/lambda.hat = 1/(μ/σ2) = 1/0.006474599
= 154.449730 feet
Shape = alpha.hat = μ * λ = 3.402475 ft^2



X - statistics:

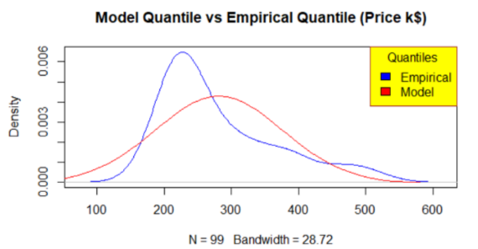
μ = 936.2213 ft
σ2 = 81165.08 ft^2
Scale = 1/lambda.hat = 1/ (μ/σ2) = 1/ 0.01153478
= 86.69433 feet
Shape = alpha.hat = μ*λ = 10.79911 ft^2

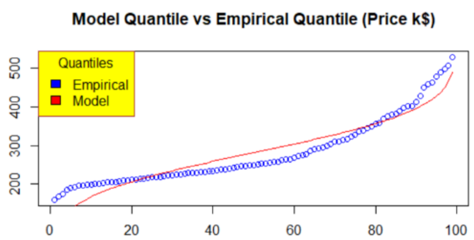


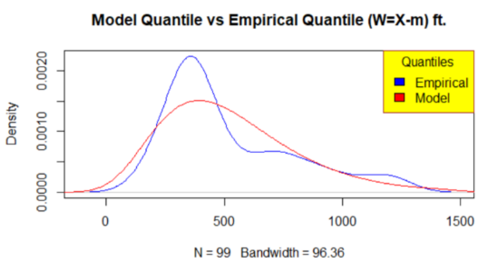
Model Percentile vs. Empirical Percentile

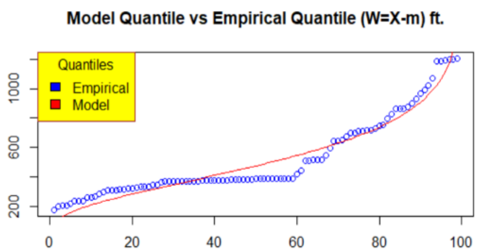
p / Perc.		
Price $ Y	**Model Perc.** qnorm(p,mean,sd)	**Empirical Perc.**
10%	167,175.2	198,969.2
50%	281,171.9	249,075.7
75%	341,169.2	326,964.9
90%	395,168.6	411,702.2
Area ft X	**Model Perc.** qgamma(p, α,1/λ)	**Empirical Perc.**
10%	594.6683	670.890
50%	907.4872	798.280
75%	1109.505	1121.950
90%	1314.825	1378.806
W = X-m	**Model Perc.** qgamma(p, α,1/λ)	**Empirical Perc.**
10%	209.1304	260.180
50%	475.022	387.570
75%	679.9692	711.240
90%	907.5268	968.096

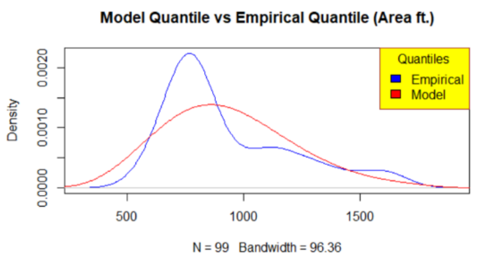
From the table above, we can see that the model and the empirical percentiles are not similar. Despite of that, the graphs below show that the models fit the data.

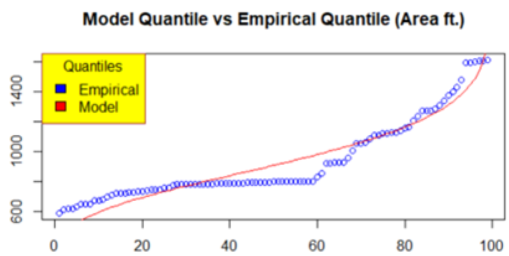












Confidence interval

Y variable statistics:

----

n=267

$\bar{X}$=mean.y=281,171.9$

S^2=var.y= 7942217729$^2

S=sd.y=89,119.12$

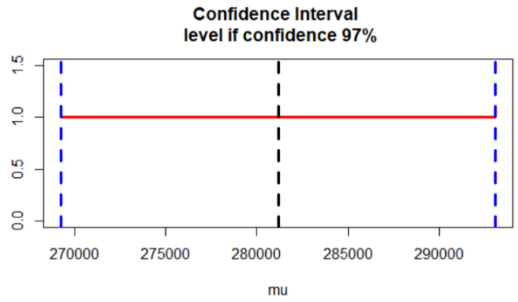
I. Confidence interval for the mean with Unknown variance:

level of confidence = 97%

1-α/2= 1-(1-0.97)/2=0.985

t-distribution:

qt(0.985,267-1)= 2.181796



$$\left[\bar{X}_n - \frac{S}{\sqrt{n}} t_{n-1,1-\alpha/2} \, , \, \bar{X}_n + \frac{S}{\sqrt{n}} t_{n-1,1-\alpha/2}\right]$$

interval confidence = [269272.4, 293071.4]

Length of interval = 23799.03

II. Confidence interval for the variance:

Level of confidence = 92%

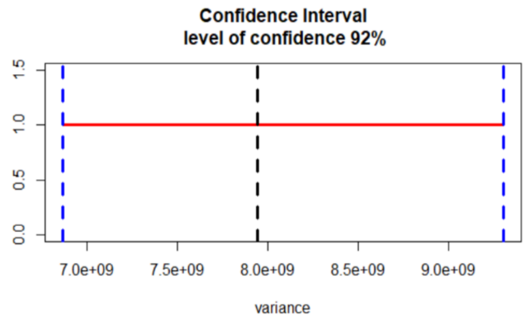
n-1=267-1=266

α/2= 0.04

qchisq(0.04,266)= 227.0293

1-α/2= 1-(1-0.92)/2=0.96

qchisq(0.96,266)= 307.7226



$$\left[\frac{(n-1)S_n^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} , \frac{(n-1)S_n^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right]$$

Interval confidence for variance= [6865371170, 9305536861]

Length of interval = 2440165691

Hypothesis testing

The data is divided into two groups:

p0: $\{(Area\ \boldsymbol{i}, Price\ \ \boldsymbol{i}) \,|\, Area\ \boldsymbol{i} <$ median$(Area)\}$

p1: $\{(Area\ \boldsymbol{i}, Price\ \ \boldsymbol{i}) \,|\, Area\ \boldsymbol{i} >=$ median$(Area)\}$

where $\boldsymbol{i}$ is the index and the median($Area$) = 798.28 feet.

Now, we are going to test the null and the alternative hypotheses:

H0:the mean in dollars of the housing prices where the area is greater or equal to the median 798.28 ft **is equal to** the mean in dollars of the housing prices where the area is smaller than the median 798.28 ft.

H1: the mean in dollars of the housing prices where the area is greater or equal to the median 798.28 ft **is not equal to** the mean in dollars of the housing prices where the area is smaller than the median 798.28 ft.

Statistically, the hypotheses can be written as follows:

$$H_0 : \mu_{y.p_1} = \mu_{y.p_0}$$
$$H_1 : \mu_{y.p_1} \neq \mu_{y.p_0}$$

The variances of the samples are unknown and inequal. The test statistic under the null hypothesis can be approximated by a normal distribution

$$T = \frac{\bar{X}_m - \bar{Y}_n}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}} \overset{H_0}{\sim} N(0,1)$$

$\bar{X}$= Mean y.p1 =342831.7
$m =$ 134
S_x^2=7442008725

$\bar{Y}$= Mean y.p0 =219048.5071
$n =$ 133
S_y^2=758296836

T = $(342831.7 - 219048.5071)$ / sqrt (7442008725/134 + 758296836/133)= 15.81787
significance level = 3%
1-α= 1-0.03= 0.97

P-value = P_{H_0}(T > 15.81787) = 1-pnorm(15.81787) = 0.000
z1−α = qnorm(0.97)= 1.880794

- ✓ We reject null hypothesis when $\{|T| > Z_{1-\alpha}\}$
  ⇨ 15.81787 > 1.880794

- ✓ We reject null hypothesis when $\{p-value < \alpha\}$
  ⇨ 0.000 < 0.03

Inequality exists, so we will reject H0.

We conclude that there is sufficient evidence to say that the mean price of houses between these two groups (p0, p1) is not equal.
In other words, the area of the house affects the housing price, so we can say that there is a relationship between the two variables (positive relationship as we saw previously in the linear regression model).