Report

Inferential Statistics and Multivariate Regression Modeling of Housing Prices in Stockholm

Name	Date of Birth	Email
Abhay Singh	1989-11-03	abhaysingh89@hotmail.com

Using a significance level of 1%, testing the hypothesis that half of the housing units in the region have a balcony.

We want to know whether the proportion of houses having a balcony is 50%.

Hypothesis

$$H_0 : p = 0.5$$
$$H_1 : p \neq 0.5$$

-**Model Independence Assumption**: It's mentioned that Real Estate Agency collected data randomly.
-**10% Condition**: 647 housing units are less than 10% of total apartments in Stockholm County.
-**Success / Failure Condition**: Both
np $= 647(0.5) = 323.5$ and
nq $= 647(0.5) = 323.5$ are greater than 10; the sample is large enough.

The conditions are satisfied, so we can use a Normal model and perform a one-proportion z-test.

The null model is a Normal distribution with a mean of 0.5 and a standard deviation of

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.5)(1-0.5)}{647}}$$
$$= 0.01965703$$

The observed proportion is $\hat{p} = 378/647 = 0.5842349$, so

$$z = \frac{\hat{p}_n - p_0}{SD(\hat{p})} = \frac{0.5842349 - 0.5}{0.019657} = 4.2852322$$

$$P = 2.P(z \geq 4.2852322) = 1.8254844 \times 10^{-5}$$

The P-value of 1.8254844×10^{-5} indicates that, if the true proportion of houses with balconies was 50%, the probability of having observed sample proportion of 0.5842349 is nearly zero. Therefore, we reject the null hypothesis (H_0). This provides strong evidence that the proportion of houses with a balcony is not 50%.

Also, the Confidence Interval is (0.5343255 , 0.6341444). With 99% confidence, the true proportion of housing units with balconies lies between 53.4% and 63.4%. The null hypothesis value (p=0.5) does not fall within this interval, further supporting the rejection of H_0.

```r
prop.test(x= has_bal, n=n, p=0.5, alternative = "two.sided", conf.level = 0.99)
```

```
    1-sample proportions test with continuity correction

data:  has_bal out of n, null probability 0.5
X-squared = 18.028, df = 1, p-value = 2.177e-05
alternative hypothesis: true p is not equal to 0.5
99 percent confidence interval:
 0.5329390 0.6337944
sample estimates:
        p
0.5842349
```

Using a significance level of 5%, testing the hypothesis that the expected size of the housing units in the region is 75 m^2.

We want to know whether average size of housing unit is 75 m^2.

Hypothesis

$$H_0 : \text{Mean area, } \mu = 75$$
$$H_1 : \text{Mean area, } \mu \neq 75$$

From the data,

$$n = 647 \text{ houses}$$
$$\bar{y} = 76.9655487 \ m^2$$
$$s = 36.2948669 \ m^2$$
$$SE(\bar{y}) = \frac{s}{\sqrt{n}} = \frac{36.2948669}{\sqrt{647}} = 1.4268985 m^2$$

$$t_{obs} = \frac{\bar{y} - \mu}{SE(\bar{y})} = \frac{76.9655487 - 75}{1.4268985} = 1.3774972 \text{ with 646 df.}$$

t_{crit} (n = 647, $\alpha = 0.05$) is in between 1.966 and 1.962 which is more than $t_{obs} = 1.377$, hence we accept the null hypothesis.

Additionally, the Confidence Interval is (74.1636295 , 79.7674679). The 95% confidence interval for the true mean is between 74.1636295 m^2 and 79.7674679 m^2. Since 75 m^2 falls within the confidence interval, this also suggests that the sample mean is not significantly different from 75 m^2 at the 5% significance level.

```
t.test(DataSet$AREA, mu = 75, alternative = "two.sided", conf.level = 0.95)
```

```
    One Sample t-test

data:  DataSet$AREA
t = 1.3775, df = 646, p-value = 0.1688
alternative hypothesis: true mean is not equal to 75
95 percent confidence interval:
 74.16363 79.76747
sample estimates:
mean of x
 76.96555
```

Estimating the proportion of apartments in the region through a 90% confidence interval.

The observed proportion is $\hat{p} = 518/647 = 0.8006182$, so

$$SD(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.8006182)(1-0.8006182)}{647}}$$
$$= 0.0157074$$

Z-critical value for the 90% confidence level is 1.6448536.

$$\begin{aligned}\text{CI} =& \hat{p} \pm Z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ =& 0.8006182 \pm (1.6448536).(0.0157074) \\ =& (0.7747819, 0.8264546)\end{aligned}$$

With 90% confidence we can say that proportion of apartment lies between 0.7747819 and 0.8264546.

```
prop.test(sum(DataSet$TYPE == "Apartment"), n = n, conf.level = 0.90)
```

```
    1-sample proportions test with continuity correction

data:  sum(DataSet$TYPE == "Apartment") out of n, null probability 0.5
X-squared = 232.68, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
90 percent confidence interval:
 0.7727467 0.8259114
sample estimates:
        p
0.8006182
```

Estimating the expected number of rooms of the housing units in the region through a 95% confidence interval.

$$n = 647, \quad df = 646$$
$$\bar{y} = 3.0540958, \quad s = 1.4515427$$
$$SE(\bar{y}) = \frac{s}{\sqrt{n}} = \frac{1.4515427}{\sqrt{647}} = 0.057066$$

t-critical value with 646 degree of freedom for the 95% confidence level is 1.963643

$$\begin{aligned} \text{CI} =& \bar{y} \pm t_{crit} \cdot SE(\bar{y}) \\ =& 3.0540958 \pm (1.963643).(0.057066) \\ =& (2.9420385, 3.1661531) \end{aligned}$$

With 95% confidence we can say that the expected numbers of rooms lies between 2.9420385 and 3.1661531.

```r
C.I_4 = t.test(DataSet$ROOMS,conf.level=0.95)
C.I_4$conf.int
```

```
[1] 2.942039 3.166153
attr(,"conf.level")
[1] 0.95
```

Fitting a linear regression that explains STARTING_PRICE in terms of some or all of the remaining variables.

We need to study how the starting price (y_i) is affected by the remaining variables (x_1, x_2, …). The model will then be:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \epsilon_i \quad with \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

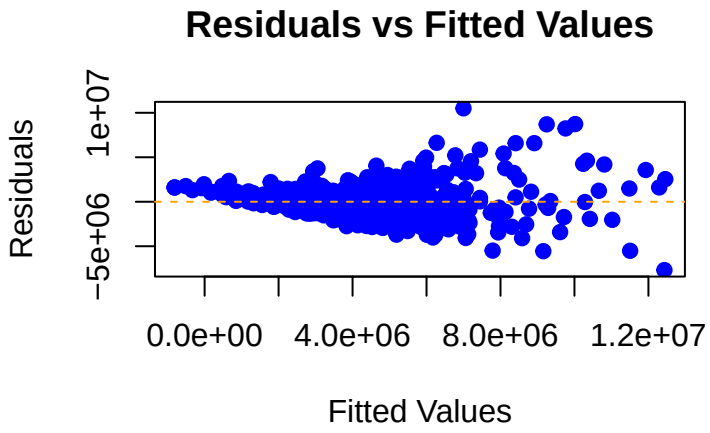
```
mult_reg = lm(STARTING_PRICE~REGION+TYPE+BALCONY+ROOMS+AREA,data=DataSet)
summary(mult_reg)
```

```
Call:
lm(formula = STARTING_PRICE ~ REGION + TYPE + BALCONY + ROOMS +
    AREA, data = DataSet)

Residuals:
     Min       1Q   Median       3Q      Max
-7682432 -1108300   -22441   877784 10500686

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -176230     287731  -0.612 0.540437
REGIONNorthwest   395799     294202   1.345 0.178997
REGIONSoutheast  -920040     282194  -3.260 0.001172 **
REGIONStockholm  1510987     210209   7.188 1.85e-12 ***
REGIONWest      -1501363     262732  -5.714 1.69e-08 ***
TYPETerrace     -1172563     340028  -3.448 0.000601 ***
TYPEVilla         480242     332675   1.444 0.149350
BALCONYYes       -257932     170062  -1.517 0.129839
ROOMS            -160632     126901  -1.266 0.206044
AREA               60419       4833  12.500  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1835000 on 637 degrees of freedom
Multiple R-squared:  0.5628,    Adjusted R-squared:  0.5566
F-statistic:  91.1 on 9 and 637 DF,  p-value: < 2.2e-16
```

The Multiple regression between Region, Type, Balcony, Rooms and Area as a predictor of Starting Price is obtained as:

$$\begin{aligned}\text{Starting Price} = & -176230 + 395799 \cdot \text{REGIONNorthwest} - 920040 \cdot \text{REGIONSoutheast} \\ & + 1510987 \cdot \text{REGIONStockholm} - 1501363 \cdot \text{REGIONWest} - 1172563 \cdot \text{TYPETerrace} \\ & + 480242 \cdot \text{TYPEVilla} - 257932 \cdot \text{BALCONYYes} - 160632 \cdot \text{ROOMS} + 60419 \cdot \text{AREA}. \end{aligned} \tag{1}$$

This equation yields R-square as 0.5628. We can improve the model by transforming one or more variables and including interaction between variables.



The residuals vs fitted graph gives us a funnel shaped pattern. Thus log transformation might yield better results.

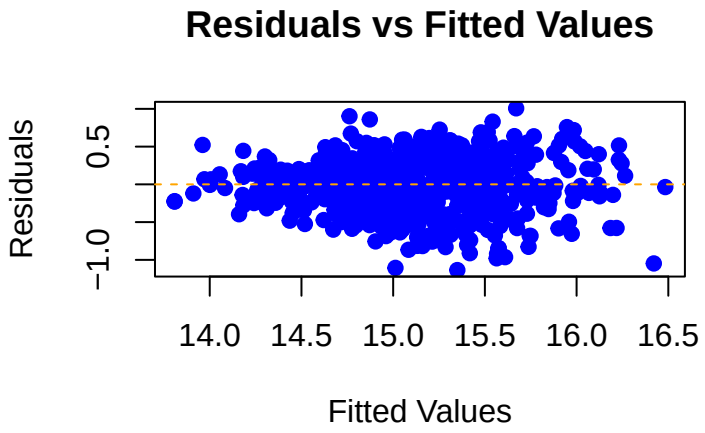
```
mult_reg2 = lm(log(STARTING_PRICE)~REGION+TYPE+ROOMS+BALCONY+AREA+AREA*ROOMS,data=DataSet)
summary(mult_reg2)
```

```
Call:
lm(formula = log(STARTING_PRICE) ~ REGION + TYPE + ROOMS + BALCONY +
    AREA + AREA * ROOMS, data = DataSet)

Residuals:
     Min       1Q   Median       3Q      Max
-1.13547 -0.21670  0.02438  0.25505  1.00717

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)     13.8988571  0.0801428 173.426  < 2e-16 ***
REGIONNorthwest  0.1850940  0.0566728   3.266 0.001149 **
REGIONSoutheast -0.1972418  0.0545318  -3.617 0.000322 ***
REGIONStockholm  0.3768469  0.0403820   9.332  < 2e-16 ***
REGIONWest      -0.4179828  0.0507025  -8.244 9.56e-16 ***
TYPETerrace     -0.0762431  0.0653602  -1.167 0.243847
TYPEVilla        0.1829979  0.0657026   2.785 0.005508 **
ROOMS            0.1103812  0.0284575   3.879 0.000116 ***
BALCONYYes      -0.0697553  0.0330916  -2.108 0.035426 *
AREA             0.0138216  0.0014257   9.695  < 2e-16 ***
ROOMS:AREA      -0.0010429  0.0002348  -4.442 1.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3525 on 636 degrees of freedom
Multiple R-squared:  0.6113,    Adjusted R-squared:  0.6052
F-statistic:   100 on 10 and 636 DF,  p-value: < 2.2e-16
```

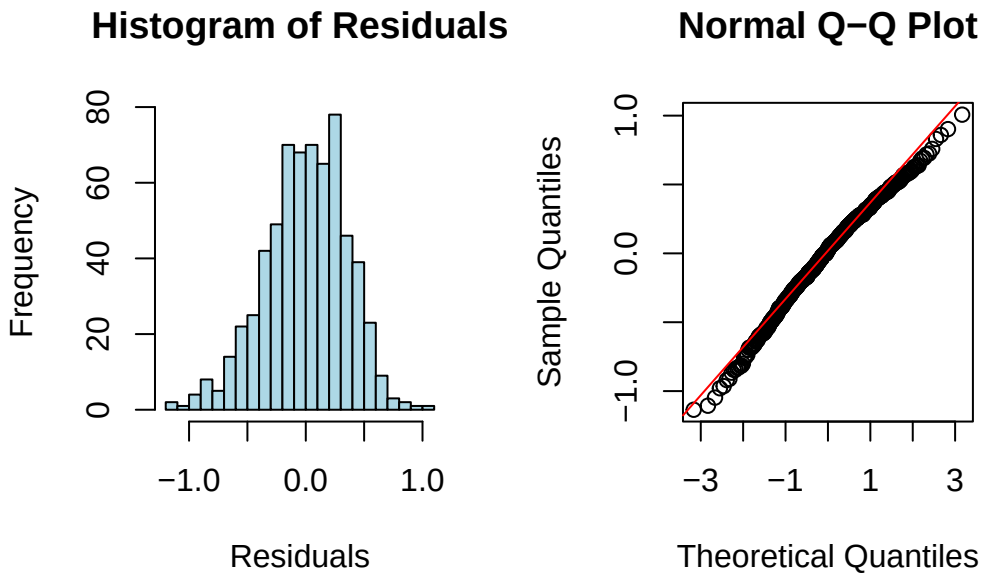
The modified equation is

$$\begin{aligned}\log(\text{STARTING\_PRICE}) = 13.899 &+ 0.185 \cdot \text{REGIONNorthwest} - 0.197 \cdot \text{REGIONSoutheast} \\ &+ 0.377 \cdot \text{REGIONStockholm} - 0.418 \cdot \text{REGIONWest} - 0.076 \cdot \text{TYPETerrace} + 0.183 \cdot \text{TYPEVilla} \\ &+ 0.110 \cdot \text{ROOMS} - 0.070 \cdot \text{BALCONYYes} + 0.014 \cdot \text{AREA} \\ &- 0.001 \cdot (\text{ROOMS} \times \text{AREA}). \end{aligned} \tag{2}$$



The Fitted vs. Residuals plot shows that the points are randomly scattered without any systematic patterns, and the variation around the estimated regression appears mostly constant.

Therefore, the assumptions of linearity and variance homogeneity (constant variance) are satisfied.



The histogram of the residuals displays a bell-shaped distribution, and the data points in the QQ-plot closely follow a straight line. This indicates that the assumption of normality of the errors is valid.

Additionally, since the data were collected randomly, we can reasonably assume that the observations are independent, and no clusters or groupings are observed in the data.

This regression model yields R-square as 0.6113.

Using a 5% significance level, $\alpha = 0.05$.

Hypothesis

H_0 : The coefficient for a variable is zero (i.e., the variable has no effect).

H_1 : The coefficient is non-zero (i.e., the variable does have an effect).

The null hypothesis is rejected if p-value < 0.05.

From the regression outputs, the statistically significant variables (with p-values < 0.05) are:

```
[1] "(Intercept)"     "REGIONNorthwest" "REGIONSoutheast" "REGIONStockholm"
[5] "REGIONWest"      "TYPEVilla"       "ROOMS"           "BALCONYYes"
[9] "AREA"            "ROOMS:AREA"
```

Predicting the starting price of these housing units in data set test.xlsx

```
test_data  = read_xlsx("test.xlsx")
test_data =data.frame(test_data)
exp(predict(mult_reg2,newdata =test_data ,interval ="prediction",level =0.95))
```

```
      fit     lwr      upr
1 3747555 1864607  7531971
2 4117685 2051409  8265212
3 6838775 3392129 13787459
4 3350357 1672540  6711286
5 2562945 1277154  5143222
6 2207375 1100005  4429530
```

The following table presents the point prediction and interval prediction with 95% confidence level.

ID	REGION	TYPE	ROOMS	AREA	BALCONY	Fit	Lower	Upper
629	Northwest	Apartment	3	74	Yes	3,747,555	1,864,607	7,531,971
718	Northeast	Apartment	4	99.5	Yes	4,117,685	2,051,409	8,265,212
1534	Stockholm	Apartment	6	115	Yes	6,838,775	3,392,129	13,787,459
1695	Stockholm	Apartment	2	45	No	3,350,357	1,672,540	6,711,286
1864	Stockholm	Apartment	1	29	No	2,562,945	1,277,154	5,143,222
2138	Northeast	Apartment	2	47.5	Yes	2,207,375	1,100,005	4,429,530