Green Collar Agritech Solutions Private Limited

Data Science internship

Submittted By:
Rahul Bhuva
9033375083

The clock prices data set contains the selling Price (in pounds Stirling) of 32 antique grandfather clocks in different auctions, along with the Age of the clocks in years and the number of Bidders participating in that auction.

VARIABLES

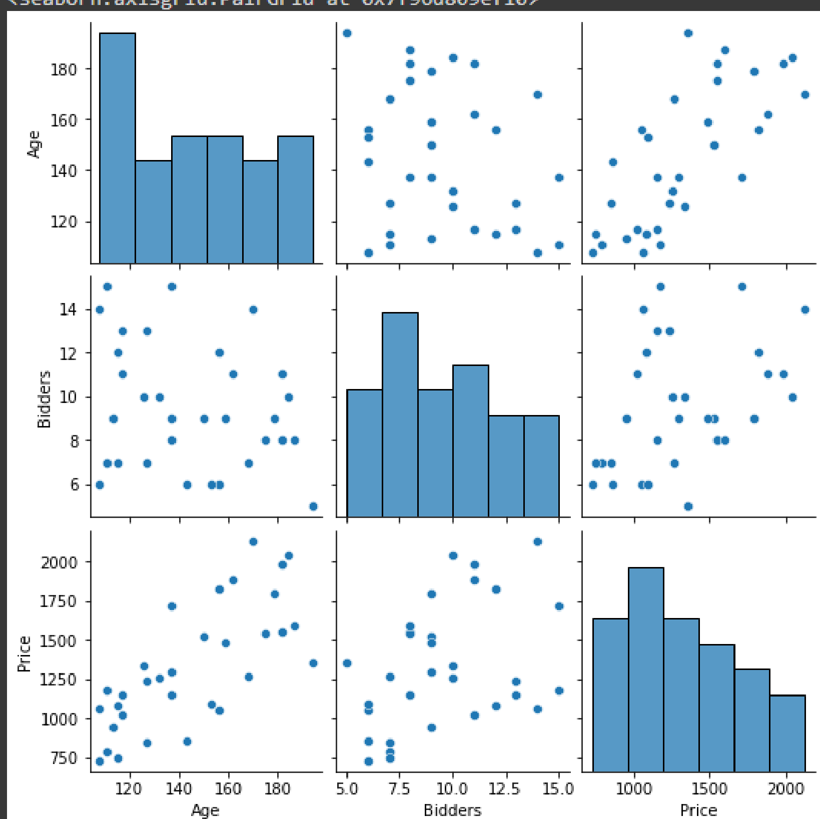
1. Age - Age of the clock (years)
2. Bidders - Number of individuals participating in the bidding
3. Price - Selling price (pounds Stirling)

We're interested in modeling the Price (Dependent Variable) based on Age and Bidders.

1. Graphically analyze the data and comment on how the age of the clock and the number of bidders are affecting the auctioned selling price.

```
sns.pairplot(df)
```

```
<seaborn.axisgrid.PairGrid at 0x7f96d809ef10>
```



Looking at the above visulization, we can say that,

1. Price of the Clock seems to be linearly related to the Age of the Clock.

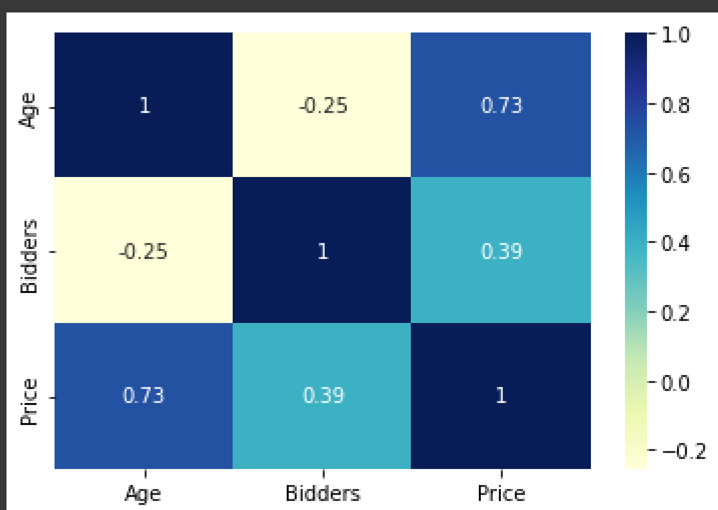
2. The Price of the Clock seems to be linearly related to the Number of Bidders on the Clock.
3. Age of the Clock and Number of Bidders don't seem to have a strong correlation between each other.

Lets look at the Correlation between the different Variables.

```
[19] print(df.corr())
```

```
              Age   Bidders     Price
Age      1.000000 -0.253749  0.730233
Bidders -0.253749  1.000000  0.394640
Price    0.730233  0.394640  1.000000
```

```
[22] dataplot = sns.heatmap(df.corr(), cmap="YlGnBu", annot=True)
     mp.show()
```



The correlation Matrix simply confirms our inferences from the visual inspection of plots.

2. Fit a first order multiple regression model to the data and answer the following based on this model :

we proceed with fitting a Full First Order Model, to explain the relationship between Price of the Clock and Age of the Clock and/or Number of Bidders for the Clock.

```python
#OLS Regression

import statsmodels.api as sm

X = df_EV[["Age", "Bidders"]]
y = df_DV["Price"]

X = sm.add_constant(X)
model = sm.OLS(y, X).fit()

# Print out the statistics
model.summary()
```

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tsa/tsatools.
  x = pd.concat(x[::order], 1)
```

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.893
Model:	OLS	Adj. R-squared:	0.885
Method:	Least Squares	F-statistic:	120.7
Date:	Mon, 22 Aug 2022	Prob (F-statistic):	8.77e-15
Time:	19:07:30	Log-Likelihood:	-200.35
No. Observations:	32	AIC:	406.7
Df Residuals:	29	BIC:	411.1
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1336.7221	173.356	-7.711	0.000	-1691.275	-982.169
Age	12.7362	0.902	14.114	0.000	10.891	14.582
Bidders	85.8151	8.706	9.857	0.000	68.010	103.620

Omnibus:	6.587	Durbin-Watson:	1.864
Prob(Omnibus):	0.037	Jarque-Bera (JB):	2.018
Skew:	0.040	Prob(JB):	0.365
Kurtosis:	1.772	Cond. No.	1.09e+03

Notes:

1. Standard Errors assume that the covariance matrix of the errors is correctly specified.
2. The condition number is large, 1.09e+03. This might indicate that there are strong multicollinearity or other numerical problems.
3. As can be observed from the R sq. value, the Model with both Age of the Clock and Number of Bidders is explaining **89.30%** of the variability in Price.

(From now onwards I am using R Software for the further analysis)

a. Is the Model useful?

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Untitled1* clock

Source on Save

```r
plot(clock.data, pch=16)

mlrm1 <- lm(Price ~ Age + Bidders, data=clock.data)
summary(mlrm1)



summary(mlrm1)
anova(mlrm1)


```

26:1 (Top Level)

Console Terminal Background Jobs

R 4.2.1 · ~/

```
> summary(mlrm1)

Call:
lm(formula = Price ~ Age + Bidders, data = clock.data)

Residuals:
   Min     1Q Median     3Q    Max 
-207.2 -117.8   16.5  102.7  213.5 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1336.7221   173.3561  -7.711 1.67e-08 ***
Age            12.7362     0.9024  14.114 1.60e-14 ***
Bidders        85.8151     8.7058   9.857 9.14e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 133.1 on 29 degrees of freedom
Multiple R-squared:  0.8927,	Adjusted R-squared:  0.8853 
F-statistic: 120.7 on 2 and 29 DF,  p-value: 8.769e-15

> anova(mlrm1)
Analysis of Variance Table

Response: Price
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Age        1 2554859 2554859 144.136 8.957e-13 ***
Bidders    1 1722301 1722301  97.166 9.135e-11 ***
Residuals 29  514035   17725                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

The low p-values observed for both $\beta0$ and $\beta1$ is quite low allowing us to conclude that both the values are significant.

We can also proceed with creation of ANOVA Table that allows us to infer if R2 obtained is significant or not.

As can be seen, p-values for both Age and Bidders is nearly 0, allowing us to conclude that both $\beta0$ and $\beta1$ are both significant.

Therefore, we can conclude that the following Model that has been fitted is useful :

Price = -1336.7221 + 12.7362(Age) + 85.8151(Bidders)

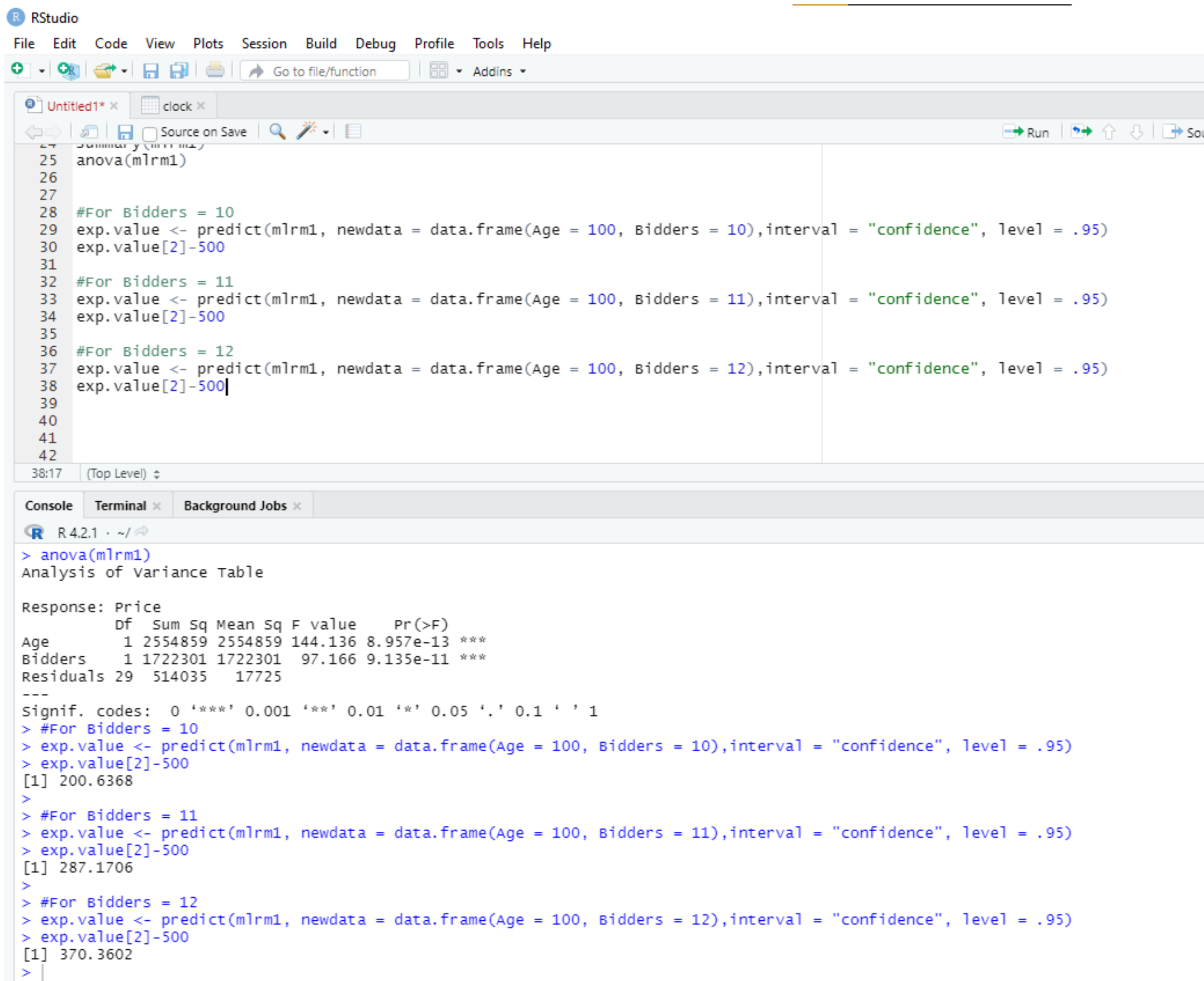
b. Given the age of a clock, by what amount can one expect the selling price to go up for one more person participating in the auction?

Using the fitted Model described above, we can say that for a Clock with given age, an increase of 1 Bidder in the number of Bidders, is associated with an increase of 85.8151 in the Mean Price of the Clock.

c. An auction house has acquired several grandfather clocks each 100 years old paying an average price of £500 per clock. From the past experience it has found that such auctions (for antique grandfather clocks) typically attract about 10-12 bidders. What can be said about its expected profit per clock with 95% confidence?

We need to find 95% Confidence Interval for the Price $ 500 for a clock that is 100 years old and has 10 Bidders.

Effectively we are finding E(Price|Age = 100, Bidders = 10), E(Price|Age = 100, Bidders = 11), and E(Price|Age = 100, Bidders = 12).



The Expected Profit per Clock that is 100 years old and has 10 Bidders, with 95% confidence is 200.6368?

The Expected Profit per Clock that is 100 years old and has 11 Bidders, with 95% confidence is 287.1706?

The Expected Profit per Clock that is 100 years old and has 10 Bidders, with 95% confidence is 370.3602?

d. You walk into an auction selling an antique 150 year old grandfather clock and find that there are 15 bidders (including yourself) participating in the auction. You are extremely keen in acquiring the clock. At least what amount should you bid for the clock, so that, you are 99% certain that nobody else can out-bid you?

For this, we need to predict a lower bound for a Predicted value of Y ($\hat{Y}$) for the given values of Age = 150 years and Bidders = 15.

```r
MSE <- mean(mlrm1$residuals^2)

beta_matrix <- as.matrix(mlrm1$coefficients, ncol=1)


new_data <- as.matrix(c(1, 150,15))

prediction_price <- t(new_data) %*% beta_matrix
t.value <- qt(0.99,30)
x <- cbind(clock.data$Age,clock.data$Bidders)
x_h <- matrix(c(150,15), nrow=1,ncol = 2)

val <- x_h %*% solve( t(x) %*% x ) %*% t(x_h)
lower.bound.price <- prediction_price-t.value*sqrt(MSE)*sqrt(1+val)
lower.bound.price
```

```
> exp.value[2]-500
[1] 287.1706
>
> #For Bidders = 12
> exp.value <- predict(mlrm1, newdata = data.frame(Age = 100, Bidders = 12),interval = "confidence", level = .95)
> exp.value[2]-500
[1] 370.3602
> MSE <- mean(mlrm1$residuals^2)
>
> beta_matrix <- as.matrix(mlrm1$coefficients, ncol=1)
>
>
> new_data <- as.matrix(c(1, 150,15))
>
> prediction_price <- t(new_data) %*% beta_matrix
> t.value <- qt(0.99,30)
> x <- cbind(clock.data$Age,clock.data$Bidders)
> x_h <- matrix(c(150,15), nrow=1,ncol = 2)
>
> val <- x_h %*% solve( t(x) %*% x ) %*% t(x_h)
> lower.bound.price <- prediction_price-t.value*sqrt(MSE)*sqrt(1+val)
> lower.bound.price
        [,1]
[1,] 1533.38
>
```

From the above calculations, we can say that if we bid at a Price higher than 1533.38, we can be 99% certain that no one else can out-bid us.

e. In presence of the other, which of the two factors, age of the clock or the number of bidders, is more important in determining the selling price of a clock?

To answer this, we first build a standardized First Order Linear Model as follows :

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

```r
std.clock <- clock.data

std.clock$Price <- (clock.data$Price - mean(clock.data$Price))/sd(clock.data$Price)
std.clock$Age <- (clock.data$Age - mean(clock.data$Age))/sd(clock.data$Age)
std.clock$Bidders <- (clock.data$Bidders - mean(clock.data$Bidders))/sd(clock.data$Bidders)

standard.model <- lm(Price ~ -1 + Age + Bidders,data = std.clock)
summary(standard.model)
vcov(standard.model)
val <- (0.88752-0.61985)/sqrt(2*(0.0038223614-0.0009699208))
p.val <- 2*(1-pt(val,30))
cat("p-value for Statistical significance between Coefficients of Age and Bidders",p.val)
```

Console Terminal Background Jobs

R 4.2.1 · ~/

```
>
> summary(standard.model)

Call:
lm(formula = Price ~ -1 + Age + Bidders, data = std.clock)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.52699 -0.29976  0.04196  0.26121  0.54305 

Coefficients:
        Estimate Std. Error t value Pr(>|t|)    
Age      0.88752    0.06183   14.36 5.60e-15 ***
Bidders  0.61985    0.06183   10.03 4.31e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.333 on 30 degrees of freedom
Multiple R-squared:  0.8927,	Adjusted R-squared:  0.8856 
F-statistic: 124.8 on 2 and 30 DF,  p-value: 2.872e-15

> vcov(standard.model)
                 Age      Bidders
Age     0.0038223614 0.0009699208
Bidders 0.0009699208 0.0038223614
>
> val <- (0.88752-0.61985)/sqrt(2*(0.0038223614-0.0009699208))
> p.val <- 2*(1-pt(val,30))
> cat("p-value for Statistical significance between Coefficients of Age and Bidders",p.val)
p-value for Statistical significance between Coefficients of Age and Bidders 0.001314188
> 
```

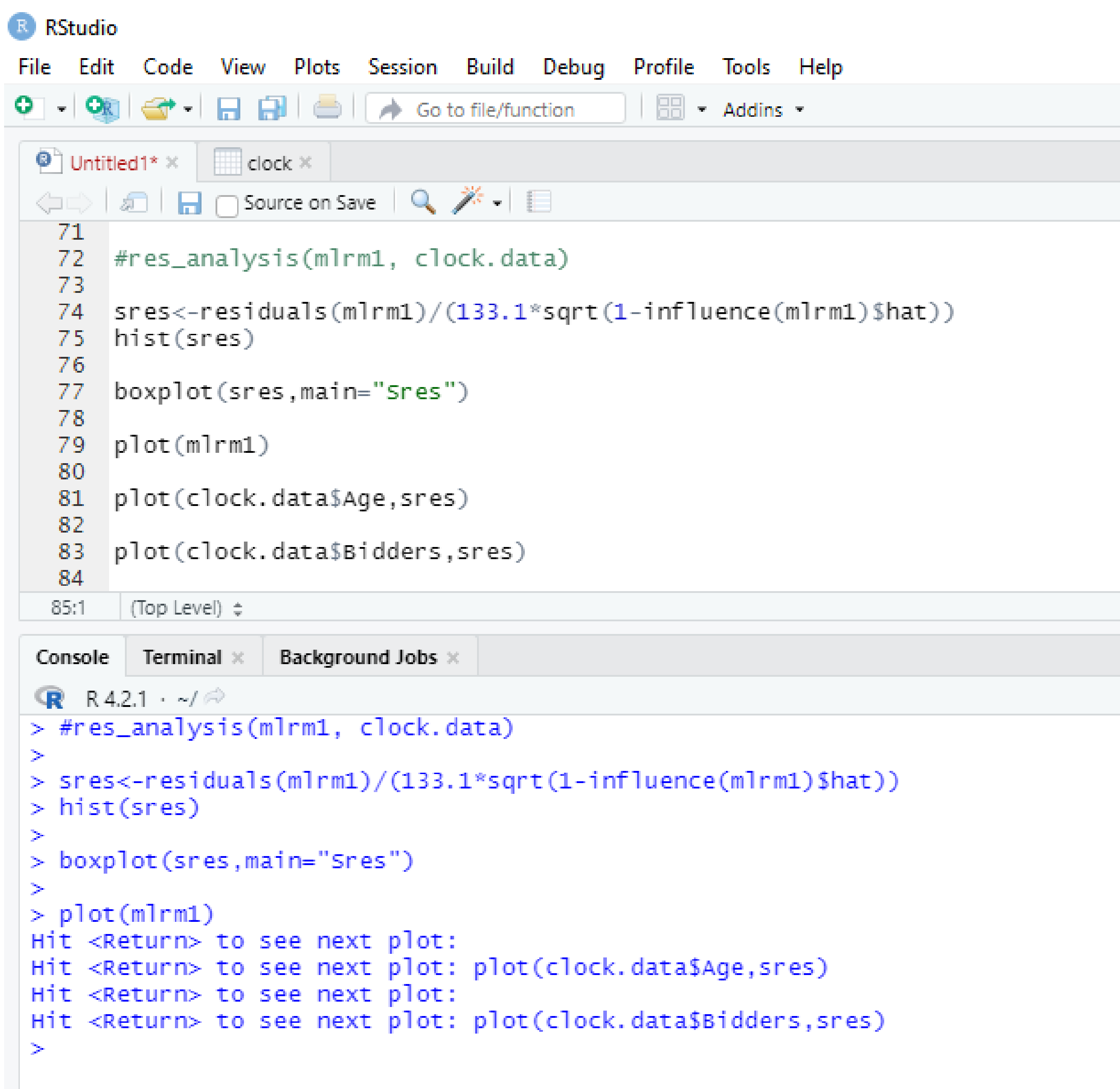
As can be observed from the calculations above, coefficient of standardized Age of the Clock is higher than the coefficient of standardized Number of Bidders for the Clock. The difference between the two values is also significant as indicated by a p-value of 0.001314. (Considering $\alpha = 5\%$)

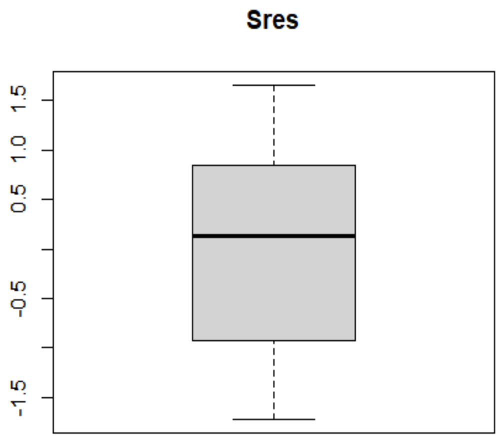
We can conclude that Age of the Clcok is more important in determining the selling price of the Clock compared to Number of Bidders.

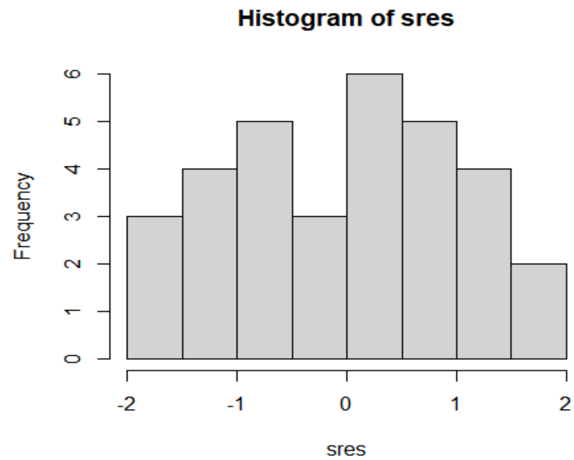
3. Is the first order model acceptable? Fit as appropriate a model as possible for the auctioned selling price of grandfather clocks, based on the information on the age of the clock and the number of bidders, and then based on this model answer the same questions as in 2. b, c, and d above.

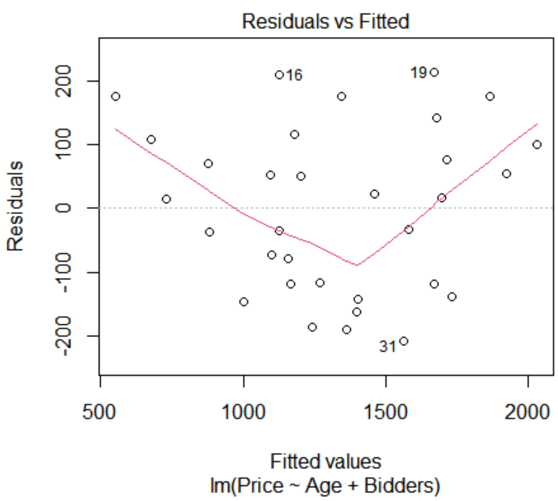
To answer whether a Model is acceptable or not, we proceed with plotting the residuals obtained after fitting the model.

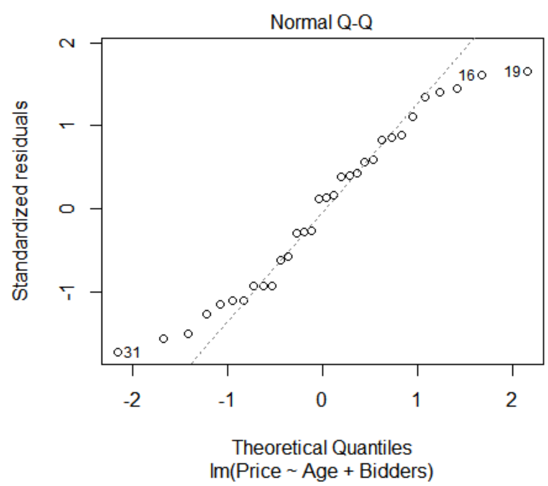
For the 1st Order Model we fitted earlier :

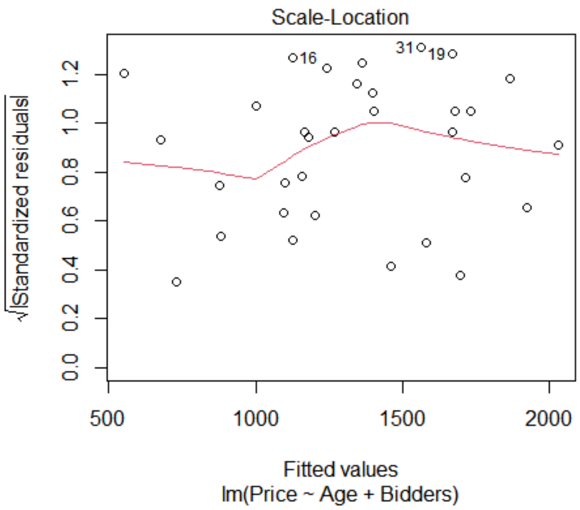


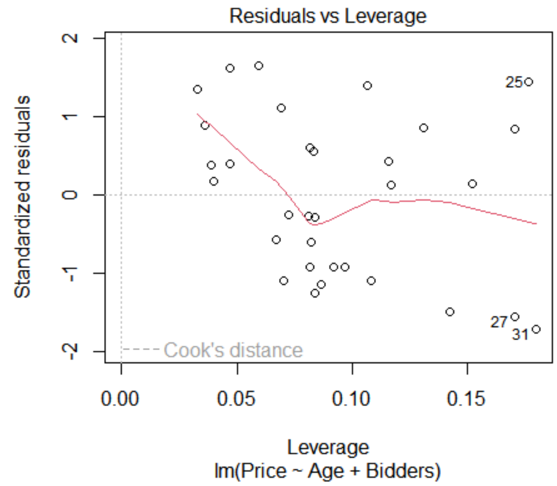

Sres
1.5
1.0
0.5
-0.5
-1.5

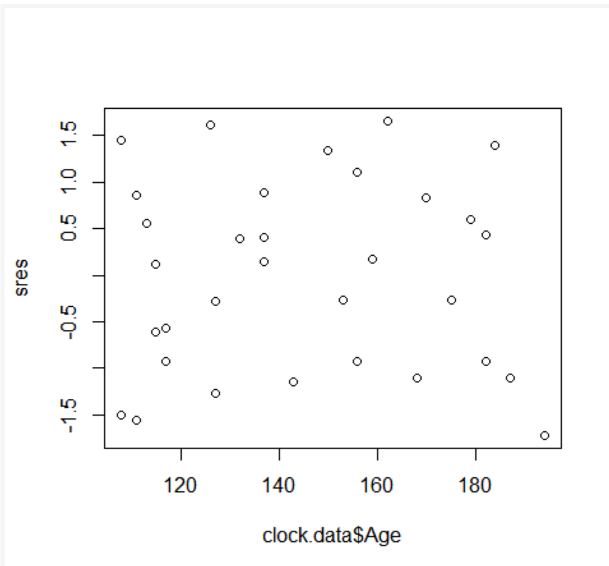

Histogram of sres
Frequency
sres

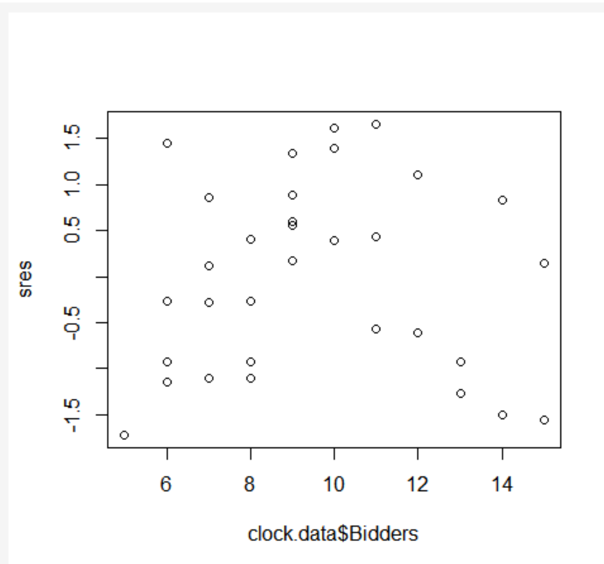

Residuals vs Fitted
Residuals
Fitted values
lm(Price ~ Age + Bidders)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Price ~ Age + Bidders)





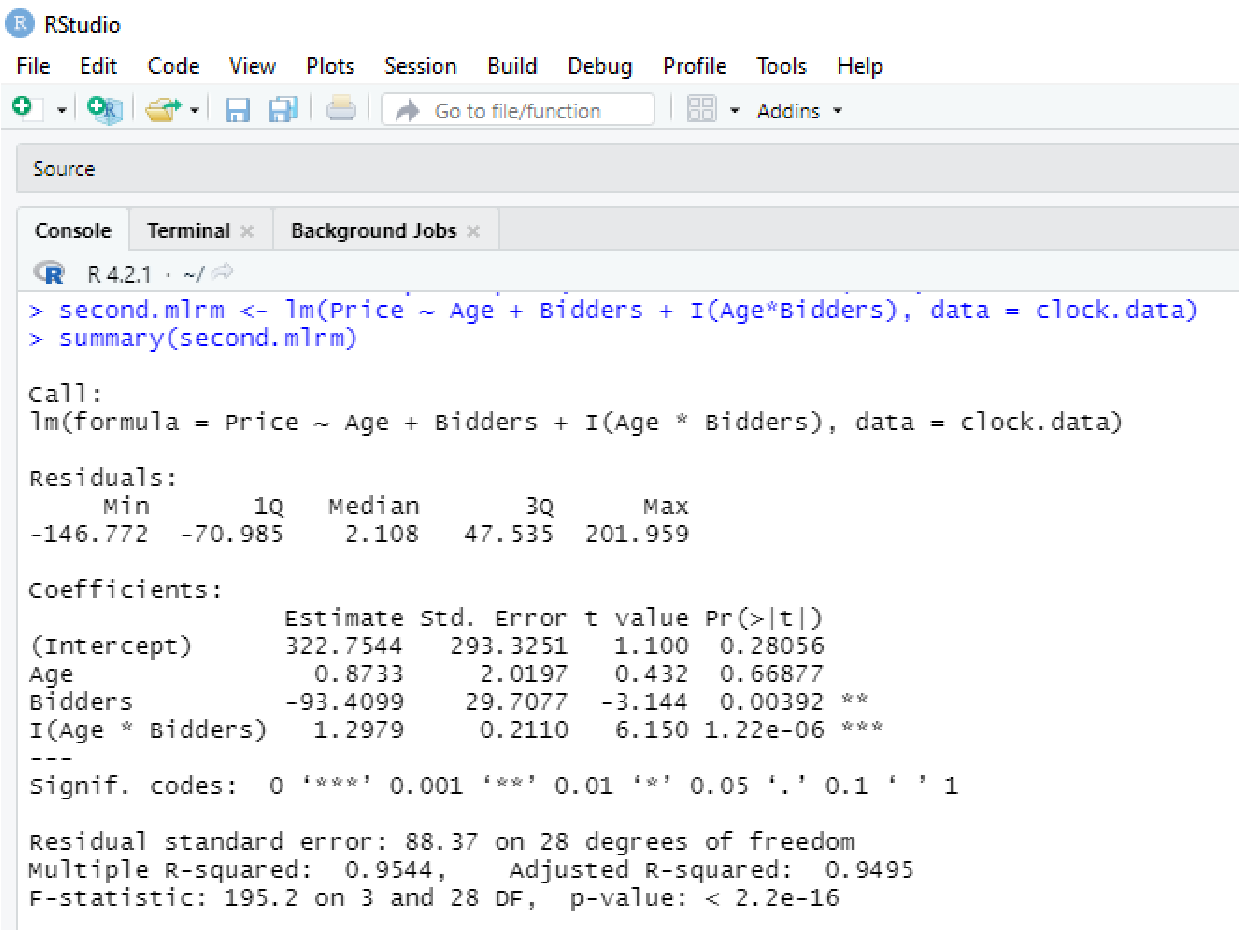




As can be seen from the Residual plots, there are no visible patterns in the Residuals obtained from the fitted model.

Based on the above, we can conclude that the First Order Model we have fitted is Acceptable.

We belive that a second order Model might be able to explain the variance in Prices better than the First Order Model already fitted. We can check this by fitting a Second Order Model as follows :




```
> second.mlrm <- lm(Price ~ Age + Bidders + I(Age*Bidders), data = clock.data)
> summary(second.mlrm)

Call:
lm(formula = Price ~ Age + Bidders + I(Age * Bidders), data = clock.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-146.772  -70.985    2.108   47.535  201.959 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       322.7544   293.3251   1.100  0.28056    
Age                 0.8733     2.0197   0.432  0.66877    
Bidders           -93.4099    29.7077  -3.144  0.00392 ** 
I(Age * Bidders)    1.2979     0.2110   6.150 1.22e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 88.37 on 28 degrees of freedom
Multiple R-squared:  0.9544,	Adjusted R-squared:  0.9495 
F-statistic: 195.2 on 3 and 28 DF,  p-value: < 2.2e-16
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Untitled1* clock

Source on Save

```
                                          NA)
))

round(tab, 2)
```

116:14 (Top Level)

Console Terminal Background Jobs

R 4.2.1 · ~/

```
>
> #fit <- lm(formula = Price ~ Age + Bidders + I(Age*Bidders),data = clock.data)
> fit.aov <- anova(second.mlrm)
> tab <- as.table(cbind(
+   'SS' = c("SSR(x1, x2, x3)" = sum(fit.aov[1:3, 2]),
+            "SSR(x1)"          = fit.aov[1, 2],
+            "SSR(x2|x1)"       = fit.aov[2, 2],
+            "SSR(x3|x1, x2)"   = fit.aov[3, 2],
+            "SSE"              = fit.aov[4, 2],
+            "Total"            = sum(fit.aov[, 2])),
+
+   'Df' = c(                     sum(fit.aov[1:3, 1]),
+                                 fit.aov[1, 1],
+                                 fit.aov[2, 1],
+                                 fit.aov[3, 1],
+                                 fit.aov[4, 1],
+                                 sum(fit.aov$Df)),
+
+   'MS' = c(                     sum(fit.aov[1:3, 2]) / sum(fit.aov[1:3, 1]),
+                                 fit.aov[1, 3],
+                                 fit.aov[2, 3],
+                                 fit.aov[3, 3],
+                                 fit.aov[4, 3],
+                                 NA)
+ ))
>
> round(tab, 2)
                        SS        Df         MS
SSR(x1, x2, x3) 4572547.99      3.00 1524182.66
SSR(x1)         2554859.01      1.00 2554859.01
SSR(x2|x1)      1722300.69      1.00 1722300.69
SSR(x3|x1, x2)   295388.28      1.00  295388.28
SSE              218646.23     28.00    7808.79
Total           4791194.22     31.00
>
```

As can be seen from the R Sq. value obtained, this Model is able to explain 95.44% of the variance in Price.

b. Given the age of a clock, by what amount can one expect the selling price to go up for one more person participating in the auction?

We can write the fitted Model as :

E(Price) = 322.7544 + 0.8733(Age) - 93.4099(Bidders) + 1.2979(Age * Bidders)

This can be re-written as :

E(Price) = 322.7544 + 0.8733(Age) + (-93.4099 + 1.2979 x Age) x Bidders

For a given Age, the Expected selling Price of a Clock will go up by (-93.4099 + 1.2979 x Age) where Age will be a constant given to us.

c. An auction house has acquired several grandfather clocks each 100 years old paying an average price of £500 per clock. From the past experience it has found that such auctions (for antique grandfather clocks) typically attract about 10-12 bidders. What can be said about its expected profit per clock with 95% confidence?

We need to find 95% Confidence Interval for the (Price - 500)? for a clock that is 100 years old and has 10 to 12 Bidders.

Effectively we are finding E(Price|Age = 100, Bidders = 10), E(Price|Age = 100, Bidders = 11), and E(Price|Age = 100, Bidders = 12).

```r
#For Bidders = 10
exp.value1 <- predict(second.mlrm, newdata = data.frame(Age = 100, Bidders = 10),interval = "confidence", level = .95)
exp.value1[2] - 500

#For Bidders = 11
exp.value2 <- predict(second.mlrm, newdata = data.frame(Age = 100, Bidders = 11),interval = "confidence", level = .95)
exp.value2[2]-500

#For Bidders = 12
exp.value3 <- predict(second.mlrm, newdata = data.frame(Age = 100, Bidders = 12),interval = "confidence", level = .95)
exp.value3[2]-500
```

132:1 (Top Level)

Console Terminal Background Jobs

R 4.2.1 · ~/

```
SSR(x2|x1)      1722300.69       1.00 1722300.69
SSR(x3|x1, x2)   295388.28       1.00  295388.28
SSE              218646.23      28.00    7808.79
Total           4791194.22      31.00
> #For Bidders = 10
> exp.value1 <- predict(second.mlrm, newdata = data.frame(Age = 100, Bidders = 10),interval = "confidence", level = .95)
> exp.value1[2] - 500
[1] 210.7254
>
> #For Bidders = 11
> exp.value2 <- predict(second.mlrm, newdata = data.frame(Age = 100, Bidders = 11),interval = "confidence", level = .95)
> exp.value2[2]-500
[1] 243.6869
>
> #For Bidders = 12
> exp.value3 <- predict(second.mlrm, newdata = data.frame(Age = 100, Bidders = 12),interval = "confidence", level = .95)
> exp.value3[2]-500
[1] 271.1562
>
```

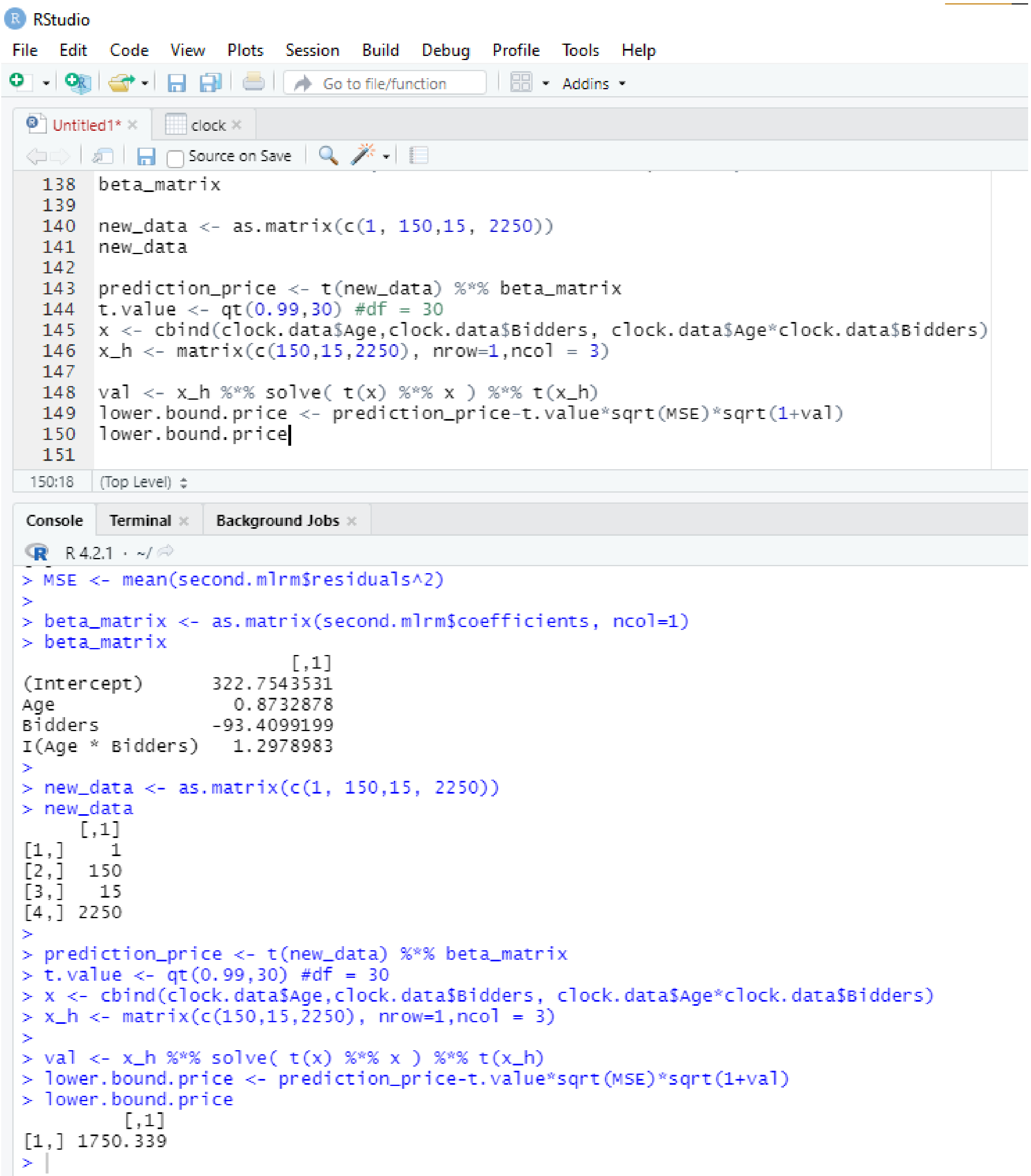
The Expected Profit per Clock that is 100 years old and has 10 Bidders, with 95% confidence is 210.7254.

The Expected Profit per Clock that is 100 years old and has 11 Bidders, with 95% confidence is 243.6869.

The Expected Profit per Clock that is 100 years old and has 12 Bidders, with 95% confidence is 271.1562.

d. You walk into an auction selling an antique 150 year old grandfather clock and find that there are 15 bidders (including yourself) participating in the auction. You are extremely keen in acquiring the clock. At least what amount should you bid for the clock, so that, you are 99% certain that nobody else can out-bid you?

For this, we need to predict a lower bound for a Predicted value of Y ($\hat{Y}$) for the given values of Age = 150 years and Bidders = 15.



From the above calculations, we can say that if we bid at a Price higher than **1750.339**, we can be 99% certain that no one else can out-bid us.