**DATA1001 Assignment 1**

**Question 1**

(a) Uncertain. First of all, it depends on the percentage of the outliers in the whole statistics say the weight of the outliers. Outliers mainly have an impact on mean and influence less on the median as the median is dominated by the range. For instance, if outliers carry little weight and indicate few deviated values in the scatter plot, refers that they are only inaccurate data. In that case, this kind of outliers can be deleted. However, if outliers are in the dominant positions of the whole data, calculation including outliers could present the deviations from the mean which is significant to the analysis. In that case, if any of the outliers are being deleted, it must have impacted the mean and the variance. Therefore, outliers should be dealt with carefully if it is not 100% for sure caused by the error.

(b) False. y and x are in the linear correlations as shown in this case, y refers to eleci and x1=educi when x2=incomei. Also, β1, β2 represent slopes which have the same sign as the covariance between y and x. As for interception, education can be both impacted by the awareness and the income. Whereas, in this model, education and income are two separate variables. Hence, β1 should be the positive number and β2 should less than zero because income is in the negative correlation with electricity. In addition, although the sign of β1β2 is certain, the influence level by either part is dominant by the coefficient.

**Question 2**

(a) The participants of the investigation are those who have already had the ability to afford the hotel prices. Thus, it is biased to the income groups such as those can't afford the price or judging the accommodation not match the value (which means that they even will not check in). Also, when choosing a certain hotel, customers prefer the categories that match their income, class, culture, food preference and so forth. Therefore, it is obvious that guests are those target customers who have the willings. Besides, there is a huge variation depends on the season and time for hotel selection and occupancy rate, which means that at a certain time the number of clusters may be limited. Thus if ' recently check-in guests' was the only picked group, there would exist a bias towards other previous or future potential guests.  Lastly, since it is a survey which at least writing and reading skills and needed handing out when check-out, it is biased to participants can read or write or understand the language that surveys provided and guests that forgetful who might forget to give back later. All the above prove that the survey handed out between current customers is biased.

(b) Firstly, it depends on how the hotel administrators define the word ‘Recently’. The types of the guests always alter with the time. For example, when approached a big festival recently there must have been a large number of customers. Thus, it biases towards other previous guests and future customers. Secondly, there excludes the group that can't access the computer since that they couldn't get the email. There is also a bias towards people may have no habits for email checking.

(c) There is a bias towards the group who are visitors since It is a hotel chain which represents guests are from all around the world not just is the residents of the cities where the hotel located. Besides, as 9 am to 4 pm is in the working hours so that some of the samples may exclude the professional class which will lead to the biases. Moreover, it also biases towards deaf or mute and low-income family, the reason is that the former may not hear or answer phones and the latter group is not accessible to the telephone which can lead to the missing of the phone-call surveys.

(d) As usual, participants who have a desire to post a review on google are those have strong willings on comments. In that case, it excludes the group not keen on reviewing and also those who don’t get accustomed to the Google review for ranking after the accommodation. Also, clusters who give review online are those already check-in the hotels. As for the hotel chain wants to expand their customers' group, they should investigate more other groups of people to satisfy their requirements. Lastly, for those can’t get access to the Internet, they are the group that is excluded.

**Question 3**

(a) 1. As median=7.50 and mean=21.73(thousands in English pounds), the plot is skewed to the right (skewed positive). The frequency decreases as the price increases.

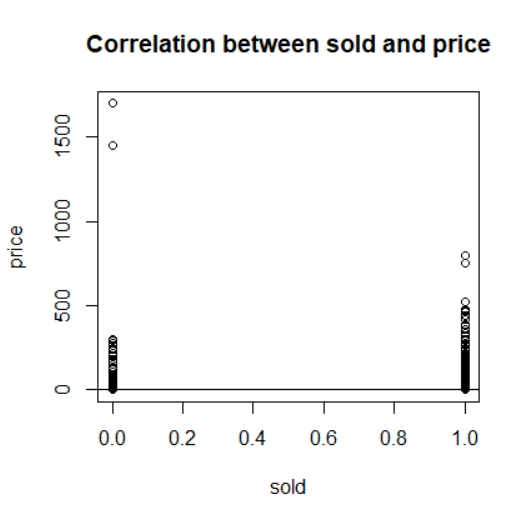
2. By applicating functions inside ‘package’, when dividing total products into ‘sold’ and ‘unsold’ categories, average highest bidding price for unsold products is 19.3 thousand pounds while for sold one, it was much higher, say 22.4 thousand pounds.

3. The minimum price of the auction is 0.03 thousand pounds and the maximum price is 1700 thousand. Thus, the range is 1699.97.

4. According to the histogram, the price range distribution mainly between 0 and 300 thousand pounds. This distribution data is unimodal.

1. yi = β0 + β1xi + ui

where: yi = price and xi =sold (while there are only two values of sold which say 1 and 0)

Variable y is explained in terms of variable x while y is the dependent variable and x is the independent variable. β1 is the slope, β0 is the intercept. ui is the disturbance/error in terms of containing unobservables.

First of all, although there are only two values in sold (1 or 0), an extremely slight positive linear relationship exists between y and x according to the coefficients as shown in the above diagram for the abline.

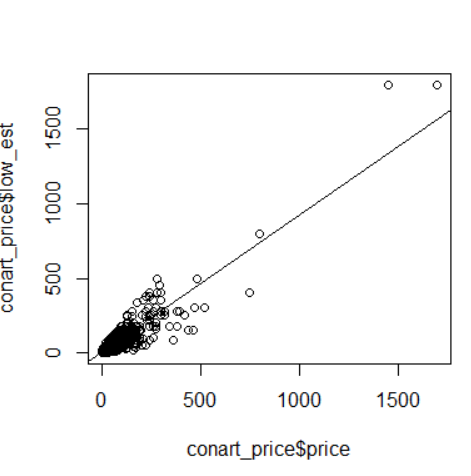
The results are going the following. The intercept, which is β0 = 0.7701502 and the gradient/slope is β1 = 0.001781. The error term is 0.4182 on 4255 degrees of freedom. This shows that the slope turns to 0 represents that the average sold price is only slightly higher than the average unsold highest bid. Also, since the intercept is around 770 English pounds, which means that for those products that are unsold (in 0), the average highest bidding is around 0.77 thousand pounds.

The price of sold products is slightly higher than the unsold one. As shown in the diagram above, the price for unsold mainly occupies in the range under 300 thousand pounds while for the other, it exists in 0-500 thousand pounds.

However, there are slight a few values that lie far off the line of best fit, as shown by the residual standard error term. And witnessed in the diagram, which is the extremely high price when sold=1(the price is above 1500 thousand pounds) and around 800 thousand pounds when sold = 1.

(c) Short report for ‘Price & low\_est’

**Relationship between price and low\_est**



According to the plot above and statistics calculated by R:

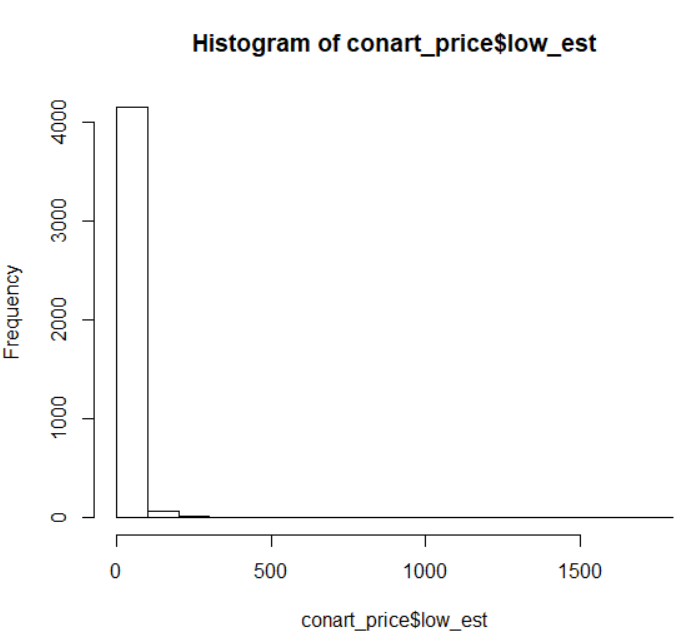
Intercept β0 = 3.6357

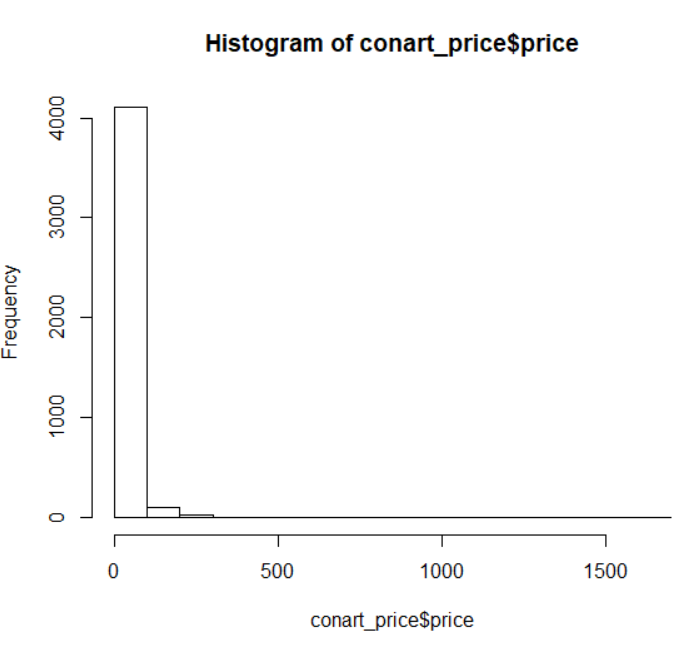
Slope β1 = 0.9182

Error value = 20.55 on 4255 degrees of freedom

These values suggest there is a positive linear correlation between the price and the estimated low\_est. As the gradient is around 0.92 which is very close to 1, it refers that the transaction price is a little higher than the estimation lowest price. However, the correlation is nowhere near perfect since the error in term of 20.55. In other words, there are a lot of outliers which are the values lie outside the line particular products value higher. At certain slots, low\_est prices are far different from the real sold one.

Besides, the intercept value explains the low\_est for products that are 0 in price (which no one participates in the auction) is around 3.63 thousand pounds.

Histograms of price and low\_est:



Summary Statistics:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| in thousands of pounds | Minimum | Maximum | Range | Median | Mean |
| price | 0.03 | 1700 | 1699.97 | 7.50 | 21.73 |
| low\_est | 0.05 | 1800 | 1799.95 | 6.00 | 19.71 |

It is clear that the distributions for both low\_est and final transaction price are skewed right, which represents that more than 50% of the auction items are less than the average price. For the extreme value no matter minimum or maximum price, low\_est always be higher indicates that expert overestimated in these two parts. Both prices for median and mean in the real transaction is bigger than low\_est refer than sold price can always reach the lowest prediction. However, real transaction price has a larger difference between mean and median (21.73-7.50=14.23) while the number for lew\_est is slightly smaller (19.71-6.00=13.71), this means of the presence of the outlier, where there is a higher average price but in the position of the lower 50th percentile. This suggests there are more prices that are low than compared to higher prices. Moreover, for the sold price, there is obviously a smaller range, but due to the higher median, suggests that a set of prices that have fewer high prices than the low\_est.

Appendix:

> conart\_price <- read.table(file=file.choose(),header=TRUE,sep="\t",dec=".")

> View(conart\_price)

> hist(conart\_price$price)

> summary(conart\_price$price)

  Min. 1st Qu.  Median Mean 3rd Qu.    Max.

  0.03    2.20 7.50   21.73 21.00 1700.00

> install.packages("dplyr")

> library(dplyr)

> conart\_price %>% group\_by(sold) %>% summarise(mean\_price=mean(price))

# A tibble: 2 x 2

sold mean\_price

<int> <dbl>

1 0 19.3

2 1 22.4

> plot(x=conart\_price$sold,y=conart\_price$price,

+ main="Correlation between sold and price",

+ xlab="sold",

+ ylab="price")

> abline(best\_fit)

>  best\_fit <- lm(conart\_price$price ~ conart\_price$sold)

>  best\_fit <- lm(conart\_price$sold ~ conart\_price$price)

> best\_fit <-lm(sold ~ price, data = conart\_price)

> best\_fit

Call:

lm(formula = sold ~ price, data = conart\_price)

Coefficients:

(Intercept)        price

 0.7701502    0.0001781

> summary(best\_fit)

Call:

lm(formula = sold ~ price, data = conart\_price)

Residuals:

   Min     1Q Median      3Q Max

-1.0728  0.2085 0.2272  0.2291 0.2298

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.7701502  0.0068899 111.780 <2e-16 \*\*\*

price       0.0001781 0.0001163   1.531 0.126

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4182 on 4255 degrees of freedom

Multiple R-squared:  0.0005507, Adjusted R-squared:  0.0003158

F-statistic: 2.344 on 1 and 4255 DF,  p-value: 0.1258

> plot(x=conart\_price$sold,y=conart\_price$price)

> abline(best\_fit)

summary(conart\_price$price)

  Min. 1st Qu.  Median Mean 3rd Qu.    Max.

  0.03    2.20 7.50   21.73 21.00 1700.00

> summary(conart\_price$low\_est)

  Min. 1st Qu.  Median Mean 3rd Qu.    Max.

  0.05    2.00 6.00   19.71 20.00 1800.00

> best\_fit <- lm(conart\_price$price ~ conart\_price$low\_est)

> best\_fit <- lm(price ~ low\_est,data=conart\_price)

> best\_fit

Call:

lm(formula = price ~ low\_est, data = conart\_price)

Coefficients:

(Intercept)      low\_est

    3.6357       0.9182

> summary(best\_fit)

Call:

lm(formula = price ~ low\_est, data = conart\_price)

Residuals:

   Min     1Q Median      3Q Max

-206.45   -3.99 -3.23   -0.16 379.07

Coefficients:

           Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.635740   0.334146 10.88 <2e-16 \*\*\*

low\_est     0.918232 0.005655  162.37 <2e-16 \*\*\*

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.55 on 4255 degrees of freedom

Multiple R-squared:  0.861, Adjusted R-squared:  0.861

F-statistic: 2.636e+04 on 1 and 4255 DF,  p-value: < 2.2e-16

> plot(x=conart\_price$price,y= conart\_price$low\_est)

> abline(best\_fit)