**Abstract.** The overall purpose is to model the relations between five explanatory variables and a response variable, and this report is made up of three sections. Firstly, we briefly introduce the problem, do exploratory analysis and give an overview of data. Secondly, we fit normal linear models to the data: verify underlying assumptions, remove outliers and refit the model. Thirdly, we interpret and evaluate our models and answer the questions. As a result, we reach a conclusion that the simple model with three explanatory variables is more appropriate.

**Introduction.** In this report, we have three objectives: summarize findings from an exploratory analysis; select a normal linear model and check its propriety for the data; and draw conclusions. We seek to find out which variable brings the largest estimated effects and whether effects of variables food, decoration and service are different according to location.

**Data.** The data consists of 168 observations of 6 variables. The response variable is the price of a meal per person, which corresponds to five explanatory variables including food, decoration, service, guide and location. Except the first three numeric random variables, guide is a categorical variable with three rating levels A, B, C and location is a binary variable where 0 represents the north and 1 represents the south.

Before generating a model, we expect to have a broad interpretation of data. Box-plots shown in Fig1 illustrate that price is not highly affected by the rating of restaurant in a guidebook: price tends to be a little higher in higher rating restaurants but difference in price is not obvious. In addition, price tends to be higher for restaurants in the south, implying that location might be a significant factor.

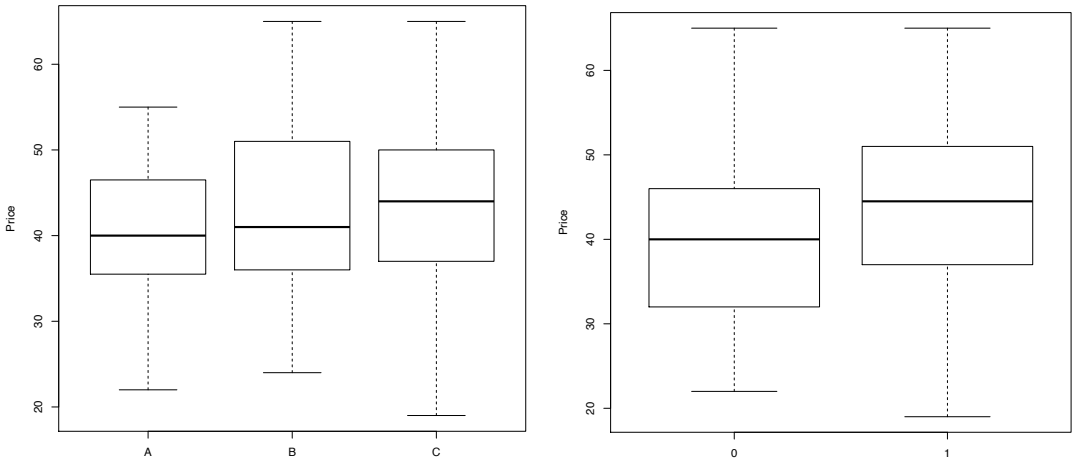


FIGURE 1. Box-plots of price of restaurants within levels of the two factors

A scatterplot shown in Fig2 agrees with the conclusion from boxplots that effect of guide is not influential, while food, decoration and service might be significant. From the plots of price against decoration, food and service, we can see that price tends to increase significantly in restaurants with higher quality of decoration, food and service. Additionally, we can see clear patterns in graphs between any two of food, decoration and service. For example, restaurants with better food tend to provide services with higher quality. Accordingly, we suspect that the effect of one variable of differs corresponding to another, thus we consider the model including the possible interaction terms.

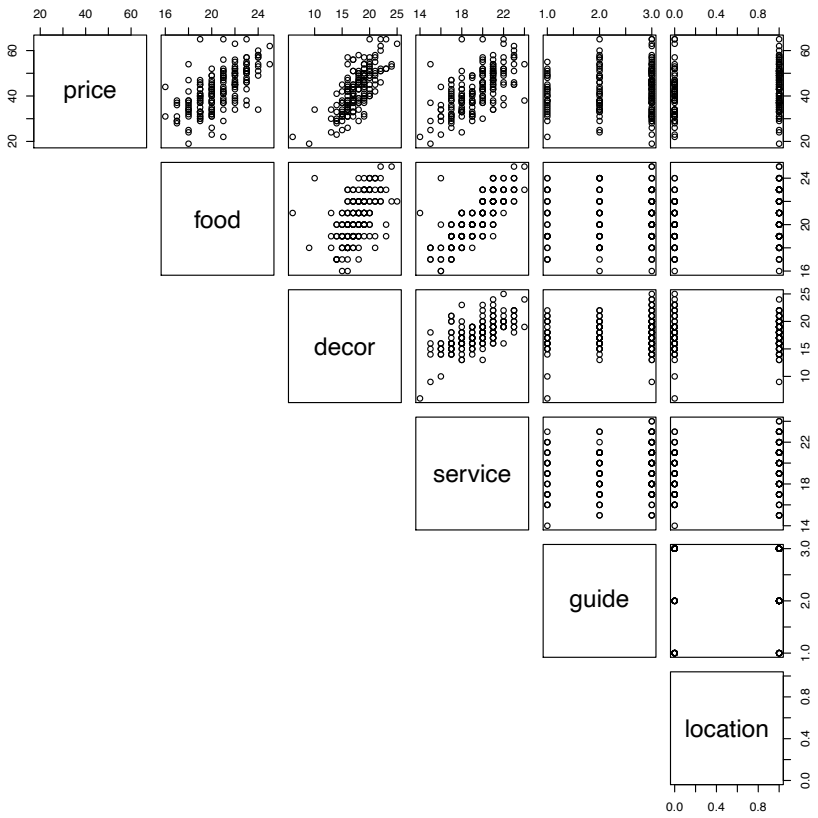


FIGURE 2. A matrix of scatterplots of six variables

**Modelling main effects.** We start with a model $LM1$ which includes the full set of variables:

$$\mathbf{E}(\boldsymbol{price}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \boldsymbol{\beta_3 x_3} + \boldsymbol{\beta_4 \mathbf{I}_{\{guide=B\}}} + \boldsymbol{\beta_5 \mathbf{I}_{\{guide=C\}}} + \boldsymbol{\beta_6 x_4}$$

This model has assumption that the error terms are independent identically distributed with constant variances. Level A of guide is the baseline. $x_1, x_2, x_3, x_4$ denote explanatory variables food, decoration, service and location respectively.

Before proceeding with the model, we check the assumptions of model and look for outliers from plots in Fig3. In (a), nearly all points have influence lower than the threshold 8/(n-2p)=8/(168-2×7)=0.0519 except for two red points which have approximately twice the Cook's distance as other points. In (b), points of studentized residuals are randomly distributed, thus the equal variance assumption is correct. And studentized residuals of nearly all data points are within the range (-3,3), and approximately 94% of 168 points are within range (-2,2), thus only a few possible

outliers exist. In (c), nearly all points have leverage lower than threshold 2p/n=2×7/168=0.0833, which is acceptable. In (d), points generally follow the straight line through the origin except on the tale of plot, thus residuals are normally distributed . From the interpretation of four plots, it is apparent that there are two outliers (red points) . Though they have leverage lower than threshold, they are highly influential, and deviates from the straight line in Q-Q plot. So we remove these two points from the data set and refit the model.

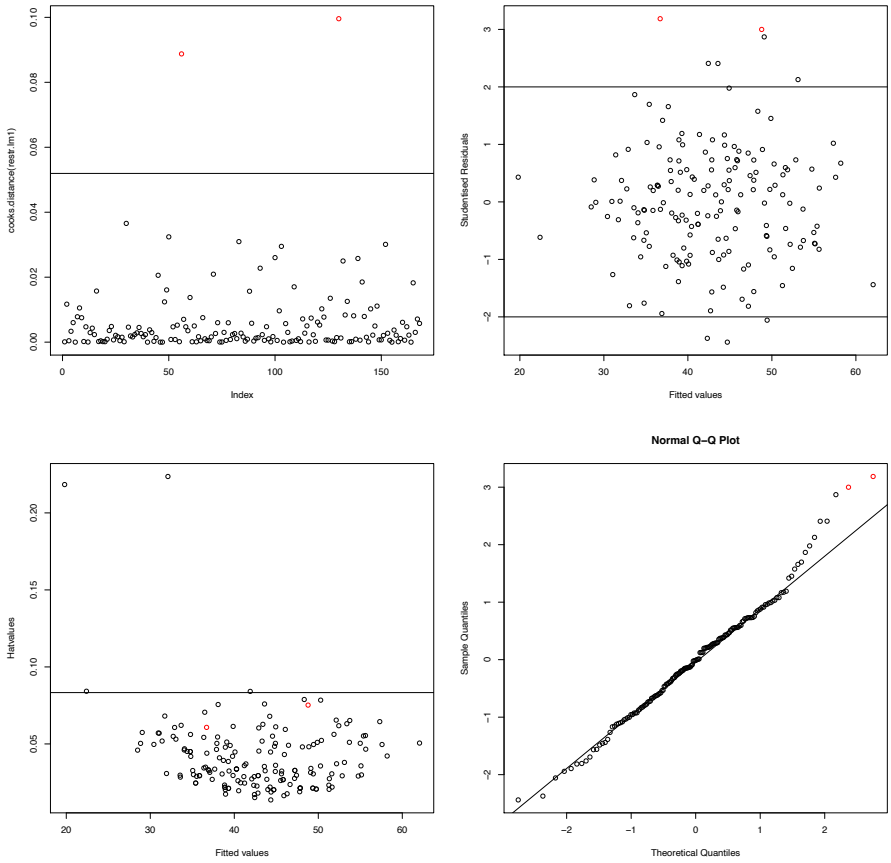


FIGURE 3. left to right, then up to down: (a) Cook's distance against price (b) Studentised residuals against fitted values (c) Hat-values against price (d) Normal Q-Q plot of studentized residuals

After refitting the model, we check if there are any outliers by calculating the Cook's distance and find that there are no points exceeding the threshold. Therefore, this model is acceptable.

From the coefficients and ANOVA table of the refitted model, it is shown that food, decoration and location are highly significant, thus we speculate that service and guide

might not be significant factors. Then we use F-test to test $H_0$: explanatory variables service and guide are not significant ($\beta_3 = \beta_4 = \beta_5 = 0$) against the general alternative. The diagram below illustrates that F statistic for dropping service and guide is 0.5054 with 3 and 159 degrees of freedom (p-value=0.6791), thus neither service nor guide are significant and should be excluded from $LM1$.

```
## Analysis of Variance Table
##
## Model 1: price ~ food + decor + location
## Model 2: price ~ food + decor + service + guide + location
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    162 4769.9
## 2    159 4724.8  3    45.055 0.5054 0.6791
```

Our model is reduced to $LM2$:

$$\mathbf{E}(\boldsymbol{price}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \boldsymbol{\beta_6 x_4}$$

To check misfit and outliers, we still use the diagnostic graphs corresponding to $LM2$ in Fig4: In (a), nearly all of points have low Cook's distance and only one red point exceeds the threshold 8/(n-2p)=8/(166-2×4)=0.0506; In (b), approximately 95% of points are within the range (-2,2); In (c), all the data points are within the threshold 2p/n=2×4/166=0.0482; In (d), points generally follow the straight Q-Q line and performs better in this model. The only possible outlier (red point) has high influence but relatively low leverage, thus it is not considered as an outlier. Therefore, this model fits the data well without any outliers.

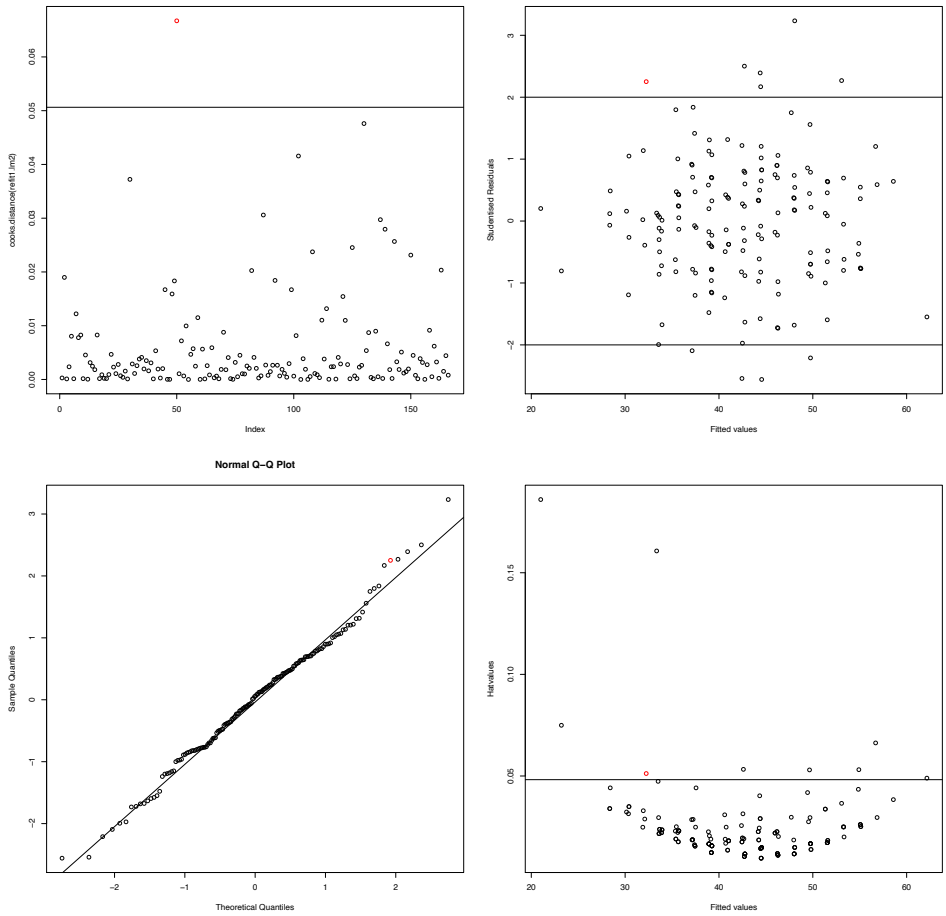


FIGURE 4. from left to right, up to down: (a) Cook's distance against price (b)Studentised residuals against fitted values (c)Hat-values against price (d)Normal Q-Q plot of studentized residuals

We can further check the significance of the three explanatory variables by T-test. For example, we test $H_0$: food is not significant ($\beta_1 = 0$) against $H_1$: food is significant. The T-statistic of the test is equal to 6.839 of an T distribution with degrees of freedom=n– p=166-4=162 with p-value =1.55e-10, which shows that food is highly significant. Similarly, decoration and location are also highly significant factors and we are unable to further simplify model $LM2$.

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26.1528     4.4628  -5.860 2.51e-08 ***
## food          1.7350     0.2537   6.839 1.55e-10 ***
## decor         1.7870     0.1843   9.696  < 2e-16 ***
## location      2.0477     0.8895   2.302   0.0226 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Modelling interaction effects.** In order to investigate the effects of interaction between location and three variables food, decoration, service, we generate a new model $LM3$ consisting of these terms:

$$\mathbf{E}(\boldsymbol{price}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \boldsymbol{\beta_3 x_3} + \boldsymbol{\beta_4}\mathbf{I}_{\{\boldsymbol{guide=B}\}} + \boldsymbol{\beta_5}\mathbf{I}_{\{\boldsymbol{guide=C}\}} + \boldsymbol{\beta_6 x_4} + \boldsymbol{\beta_7 x_1 x_4} + \boldsymbol{\beta_8 x_2 x_4} + \boldsymbol{\beta_9 x_3 x_4}$$

Performing the four plots as above, we verify the assumption that errors are independent identically normal distributed with the same variance and find four outliers. After repeating similar procedures to remove outliers and refit, we check four plots with the refitted model again with a new set of data and find no outliers. Then we stick to this model and determine whether interaction terms are significant.

To check the significance, we use a T-statistic to test the hypothesis $H_0$: interaction term of food and location is not significant ($\beta_7 = 0$) against $H_1$ that it is significant. The T-statistic is equal to 6.5107 of a T distribution with n-p=164-10=154 and p-value=0.0117, thus the interaction term of food and location is significant. Similarly, the interaction term of decoration and location is not significant as p-value = 0.4231 and interaction term of service and location is significant as p-value = 0.01579.

```
## Analysis of Variance Table
## Model 1: price ~ food + decor + service + guide + location + locatio
n:decor + location:service
## Model 2: price ~ food + decor + service + guide + location + locatio
n:food + location:decor + location:service
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    155 4656.4
## 2    154 4467.5  1    188.87 6.5107 0.0117 *
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
## Model 1: price ~ food + decor + service + guide + location + locatio
n:food + location:service
## Model 2: price ~ food + decor + service + guide + location + locatio
n:food + location:decor + location:service
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    155 4486.2
## 2    154 4467.5  1    18.717 0.6452 0.4231
```

```
## Analysis of Variance Table
## Model 1: price ~ food + decor + service + guide + location + locatio
n:food +  location:decor
## Model 2: price ~ food + decor + service + guide + location + locatio
n:food +  location:decor + location:service
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    155 4640.4
## 2    154 4467.5  1    172.83 5.9575 0.01579 *
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Interpretation.** In order to determine which model is more proper, we use an F-statistic to test the hypothesis $H_0$: $\beta_3 = \beta_4 = \beta_5 = \beta_7 = \beta_8 = \beta_9 = 0$ against $H_1$ that one of them is not 0. ANOVA table shows that F statistic is 1.7251 with corresponding p-value=0.1186, then there is no evidence against $H_0$ and we prefer the simple model $LM2$.

```
## Analysis of Variance Table
##
## Model 1: price ~ food + decor + location
## Model 2: price ~ food + decor + service + guide + location + locatio
n:food +
##     location:decor + location:service
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    160 4767.8
## 2    154 4467.5  6    300.27 1.7251 0.1186
```

In $LM2$ using the refitted data, we observe that location has the largest estimated effect on price. From the coefficients, we find that if other variables are held fixed and food is increased by 1, we would expect to see an increase of 1.735 in price. Similarly, we would expect to see an increase of 1.787 in price if decoration is increased by 1 and an increase of 2.0477 in price if location is shifted to the south. Therefore, moving from the north to the south triggers the largest estimated effect on price.

```
## Call:
## lm(formula = price ~ food + decor + location, data = refit1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.549  -3.848   0.306   3.448  16.929
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26.1528     4.4628  -5.860 2.51e-08 ***
## food          1.7350     0.2537   6.839 1.55e-10 ***
## decor         1.7870     0.1843   9.696  < 2e-16 ***
## location      2.0477     0.8895   2.302   0.0226 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.426 on 162 degrees of freedom
## Multiple R-squared:  0.6541, Adjusted R-squared:  0.6477
## F-statistic: 102.1 on 3 and 162 DF,  p-value: < 2.2e-16
```

In $LM3$, we observe that there is a positive effect between interaction food:location 2.308 and a negative effect between service:location. This illustrates that the restaurants in the south have a positive effect on the coefficient of food and a negative effect on the coefficient of service, which means that location can influence the price by changing the effects of more terms. And the location has the largest estimated effect 3.2445 compared to others which agrees with the conclusion from $LM2$.

```
## Call:
## lm(formula = price ~ food + decor + service + guide + location +
##     location:food + location:decor + location:service, data = nrefit
1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7700  -3.5940  -0.0686   3.6176  17.0526
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -27.5795     8.1070  -3.402 0.000853 ***
## food              -0.1794     0.7544  -0.238 0.812363
## decor              1.8865     0.3241   5.822 3.28e-08 ***
## service            2.0068     0.7399   2.712 0.007443 **
## guideB             0.5263     1.4272   0.369 0.712791
## guideC            -0.1697     1.2395  -0.137 0.891275
## location           3.2445     9.9487   0.326 0.744774
## food:location      2.3081     0.9046   2.552 0.011697 *
## decor:location    -0.3732     0.4646  -0.803 0.423074
## service:location  -2.1674     0.8880  -2.441 0.015789 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.386 on 154 degrees of freedom
## Multiple R-squared:  0.6639, Adjusted R-squared:  0.6443
## F-statistic:  33.8 on 9 and 154 DF,  p-value: < 2.2e-16
```

**Conclusion and evaluation.** In this report, we fit a normal linear model with all the main effects. From four plots: Cook's distance against price, studentized residuals against fitted values, hat-values against price and normal Q-Q plot of studentized residuals, we improve the model by removing outliers and refitting the model repetitively. And we reduce two least significant variables to get a smaller but fully significant model. We start a new model with interaction terms and perform similar analysis. However, the model remains relatively complex in order to test the significance of the interaction coefficients. And we reach the conclusion that there are positive interaction effects between food and location, negative effects between service and location and no effects between decoration and location. The possible drawback of modelling is that the intercepts are negative in both models which represent negative prices and are impossible in real-life situations.

**R Code.**

```
# Import data from file restr.csv
restr<-read.csv("http://www.stats.ox.ac.uk/~laws/SB1/data/restr.csv
")
attach(restr)
str(restr)

library(MASS)

# Box-plot of price affected by categorical variable guide
pdf("restrbox1.pdf",height=8,width=8)
boxplot(price~guide, xlab="Guide", ylab="Price")
dev.off()

# Box-plot of price affected by categorical variable location
pdf("restrbox2.pdf",height=8,width=8)
boxplot(price~location, xlab="Location", ylab="Price")
dev.off()

# Pairplots of six variables
pdf("restr1.pdf",height=8,width=8)
pairs(restr,lower.panel=NULL)
dev.off()

# Establish a normal linear model LM1
restr.lm1<-lm(price~food+decor+service+guide+location,data=restr)

# Four diagonostic plots to determine outliers
n1<-nrow(restr)
p1<-restr.lm1$rank
i1<-cooks.distance(restr.lm1)>(8/(n1-2*p1))

pdf("Cooks1.pdf",height=16,width=16)
par(mfrow = c(2,2))
plot(cooks.distance(restr.lm1),col=1+i1,)
abline(h=8/(n1-2*p1))

plot(rstudent(restr.lm1)~fitted(restr.lm1), xlab="Fitted values",yla
b="Studentised Residuals",col=1+i1)
abline(h=-2)
abline(h=2)

qqnorm(rstudent(restr.lm1),col=1+i1)
qqline(rstudent(restr.lm1))
```

```
plot(hatvalues(restr.lm1)~fitted(restr.lm1),xlab ="Fitted values",yl
ab="Hatvalues",col=1+i1)
abline(h=2*p1/n1)
dev.off()

# Remove outliers and refit the LM1 to new data
refit1<-restr[-which(i1), ]
refit1.lm1<-lm(price~food+decor+service+guide+location,data=refit1)
n2<-nrow(refit1)
p2<-refit1.lm1$rank
i2<-cooks.distance(refit1.lm1)>(8/(n2-2*p2))
sum(i2) # sum(i2)=0 means that there are no outliers in the refitted
model

summary(refit1.lm1)

anova(refit1.lm1)

# Compare LM1 with LM2 which drops two variables : service and guide
refit1.lm2<-lm(price~food+decor+location, data=refit1)
anova(refit1.lm2,refit1.lm1)

# Find outliers of LM2
n3<-nrow(refit1)
p3<-refit1.lm2$rank
i3<-cooks.distance(refit1.lm2)>(8/(n3-2*p3)) #Find possible outliers
 of LM1

pdf("Cooks2.pdf",height=16,width=16)
par(mfrow = c(2,2))
plot(cooks.distance(refit1.lm2),col=1+i3)
abline(h=8/(n3-2*p3))

plot(rstudent(refit1.lm2)~fitted(refit1.lm2),xlab="Fitted values",yl
ab="Studentised Residuals",col=1+i3)
abline(h=-2)
abline(h=2)

qqnorm(rstudent(refit1.lm2),col=1+i3)
qqline(rstudent(refit1.lm2))

plot(hatvalues(refit1.lm2)~fitted(refit1.lm2),xlab="Fitted values",y
lab="Hatvalues",col=1+i3)
abline(h=2*p3/n3)
dev.off()
```

```
summary(refit1.lm2)

# Establish a new model including interaction terms LM3
restr.nlm1<-lm(price~food+decor+service+guide+location+location:food
+location:decor+location:service,data=restr)
nn1<-nrow(restr)
np1<-restr.nlm1$rank
ni1<-cooks.distance(restr.nlm1)>(8/(nn1-2*np1))

pdf("NCooks1.pdf",height=16,width=16)
par(mfrow = c(2,2))
plot(cooks.distance(restr.nlm1),col=1+ni1)
abline(h=8/(nn1-2*np1))

plot(rstudent(restr.nlm1)~fitted(restr.nlm1),xlab="Fitted values",yl
ab="Studentised Residuals",col=1+ni1)
abline(h=-2)
abline(h=2)

qqnorm(rstudent(restr.nlm1),col=1+ni1)
qqline(rstudent(restr.nlm1))

plot(hatvalues(restr.nlm1)~fitted(restr.nlm1),xlab = "Fitted values
", ylab = "Hatvalues",col=1+ni1)
abline(h=2*np1/nn1)
dev.off()

# Refit the data
nrefit1<-restr[-which(ni1), ]
nrefit1.nlm1<-lm(price~food+decor+service+guide+location+location:fo
od+location:decor+location:service,data=nrefit1)
nn2<-nrow(nrefit1)
np2<-nrefit1.nlm1$rank
ni2<-cooks.distance(nrefit1.nlm1)>(8/(nn2-2*np2))
sum(ni2)

summary(nrefit1.nlm1)

anova(nrefit1.nlm1)

pdf("NCooks2.pdf",height=16,width=16)
par((mfrow = c(2,2)))

plot(cooks.distance(nrefit1.nlm1),col=1+ni2)
abline(h=8/(nn2-2*np2))
```

```
plot(rstudent(nrefit1.nlm1)~fitted(nrefit1.nlm1), xlab = "Fitted val
ues", ylab = "Studentised Residuals",col=1+ni2)
abline(h=-2)
abline(h=2)

qqnorm(rstudent(nrefit1.nlm1),col=1+ni2)
qqline(rstudent(nrefit1.nlm1))

plot(hatvalues(nrefit1.nlm1)~fitted(nrefit1.nlm1),xlab = "Fitted val
ues", ylab = "Hatvalues",col=1+ni2)
abline(h=2*np2/nn2)
dev.off()

#Generate new models to compare whether the interaction term is signi
ficant
frefit1.lm1<-lm(price~food+decor+service+guide+location+location:dec
or+location:service,data=nrefit1)
drefit1.lm1<-lm(price~food+decor+service+guide+location+location:foo
d+location:service,data=nrefit1)
srefit1.lm1<-lm(price~food+decor+service+guide+location+location:foo
d+location:decor,data=nrefit1)
anova(frefit1.lm1,nrefit1.nlm1)

anova(drefit1.lm1,nrefit1.nlm1)

anova(srefit1.lm1,nrefit1.nlm1)

#Compare LM2 and LM3 using F-test
refit1.lm3<-lm(price~food+decor+location, data=nrefit1)
anova(refit1.lm3,drefit1.lm1)
```