Practice Midterm Exam 1 Solutions

ISYE 6414

Spring Semester 2022

T/F section

Part 1 T/F

Question 1 If the confidence interval for a regression coefficient contains the value zero, we interpret that the regression coefficient is definitely equal to zero.
False. The coefficient is plausibly zero, but we cannot be certain that it is. See Topic 1.2 Lesson 4 and Lesson 5

Question 2 The larger the coefficient of determination or R-squared, the higher the variability explained by the simple linear regression model.
True. R-squared represents the proportion of total variability in Y(response) that can be explained by the regression model (that uses X). R-squared is the proportion of variability explained by the model.See Topic 1.3 Lesson 9.

Question 3 The estimators of the error term variance and of the regression coefficients are random variables.
True. The estimators are $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $\hat{\sigma^2} = \hat{\epsilon}^T \hat{\epsilon}/(n{-}p{-}1)$, where $\hat{\epsilon} = (I{-}H)Y$. These estimators are functions of the response, which is a random variable. Therefore they are also random. See Topic1.2 Lesson4.

Question 4 The one-way ANOVA is a linear regression model with one qualitative predicting variable.
True. One-way ANOVA is a linear regression model with one predicting factor/ categorical variable. See Topic 2.2 Lesson7

Question 5 We can assess the assumption of constant-variance in simple linear regression by plotting residuals against fitted values.residuals against fitted values.
True. In a residuals Vs fitted plot, if the residuals are scattered around the 0 line, it indicates that the constant variance assumption of errors hold. See Topic1.3 Lesson8

Question 6 If one confidence interval in the pairwise comparison includes zero under ANOVA, we conclude that the two corresponding means are plausibly equal.
True. If the confidence interval includes zero, it is plausible that the corresponding means are equal.See Topic2.2 Lesson5.

Question 7 In Anova, the pooled variance estimator or MSE is the variance estimator assuming equal means.
False. The pooled variance estimator is the variance estimator assuming equal variances. We assume that the variance of the response variable is the same across all populations and equal to sigma square. Module 2 Topic 2.1 Lessons 1 - 3

Question 8 Assuming the model is a good fit, the residuals in simple linear regression have constant variance.
True. Goodness of fit refers to whether the model assumptions hold, one of which is constant variance.See

Topic1.3 Lesson8

Question 9 You are interested in understanding the relationship between education level and IQ, with IQ as the response variable. In your model, you also include age. Age would be considered a controlling variable while the education level would be an explanatory variable.
True. Controlling variables can be used to control for bias selection in a sample. They're used as default variables to capture more meaningful relationships with respect to other explanatory or predicting factors. Explanatory variables can be used to explain variability in the response variable, in this case the education level.See Topic3.1 Lesson4.

Question 10 If a predicting variable is categorical with 5 categories in a linear regression model without intercept, we will include 5 dummy variables in the model.
True. When we have qualitative variables with k levels, we only include k-1 dummy variables if the regression model has an intercept. If not, we will include k dummy variables. See Topic3.1 Lesson2

Question 11 In ANOVA, the number of degrees of freedom of the chi-squared distribution for the variance estimator (not pooled variance estimator)is N-k-1 where k is the number of groups.
False. This variance estimator has N-1 degrees of freedom. We lose one DF because we calculate one mean and hence its N-1. See Topic2.1 Lesson4

Question 12 The only assumptions for a simple linear regression model are linearity, constant variance, and normality.
False The assumptions of simple Linear Regression are Linearity, Constant Variance assumption, Independence and normality. See Topic1.1 Lesson2.

Question 13 In simple linear regression, the confidence interval of the response increases as the distance between the predictor value and the mean value of the predictors decreases.
False: The confidence interval bands increase as a predictor increases in distance from the mean of the predictors. See Topic1.2 Lesson6.

Part 2 Multiple Choice

Problem 1

You are thinking about starting a new business. However, your initial capital is limited. To start, you are thinking about a pizza business but your are open to explore options with lower initial investments. For this, you collected the following data on initial investment for several types of industries:

Pizza	Bakery	Shoes	Gifts	Pets
80	150	48	100	25
125	40	35	96	80
35	120	95	35	30
58	75	45	99	35
110	160	75	75	30
140	60	115	150	28
97	45	42	45	20
50	100	78	100	75
65	86	65	120	48

Pizza	Bakery	Shoes	Gifts	Pets
79	87	125	50	20

Consider the following (incomplete) ANOVA table.

Source	Df	Sum of Squares	Mean Squares	F-statistics	p-value
Treatments	A	18186	B	C	0.00662
Error	D	E	1114		
Total	F	68336			

Question 14
What is the value for A in the ANOVA table?
☑ 4
□ 9
□ 45
□ 49

If k represents the number of levels of the qualitative variable (here k=5), this is k-1.

Question 15 What is the value for B in the ANOVA table?
□ 16
□ 371
☑ 4546
□ 50150
This is Sum of Squares Treatments / Df Treatments (i.e. 18186/4).

Question 16 What is the value for C in the ANOVA table?
□ 0.25
□ 4.00
☑ 4.08
□ 11.03
This is Mean Squares Treatment / Mean Squares Error = 4546/1114

Question 17 What is the value for D in the ANOVA table?
□ 4
□ 9
☑ 45
□ 49
If k represents the number of levels of the qualitative variable (here k=5) and N the number of observations (here N=5*10=50), this is N-k.

Question 18 What is the value for E in the ANOVA table?
□ 4456
□ 6980
☑ 50150
□ 56130

This is Sum of Squares Total − Sum of Squares Treatments = 68336 − 18186. Alternatively, this can be solved similar to B in Q15 using 1114D = 111445 (though it doesn't exactly match the solution value due to rounding).

Question 19 What is the value for F in the ANOVA table?
□ 4
□ 9
□ 45
☑ 49
If N the number of observations (here N=5*10=50), this is N-1.

Question 20 What are the null and alternative hypotheses?
□ Null: the mean initial capital is the same for all industries; Alternative: the mean initial capital is different for all industries
☑ Null: the mean initial capital is the same for all industries; Alternative: at least two industries have unequal mean initial capital
□ Null: the mean initial capital is equal to 0 for all industries; Alternative: the mean initial capital is nonzero for all industries
□ Null: the mean initial capital is equal to 0 for all industries; Alternative: the mean initial capital is nonzero for at least one industry
See Topic2.1 Lesson4

Question 21 Should we reject the null hypothesis? What are the implications in terms of the business problem?
□ Yes, we should reject the null hypothesis. You can be indifferent when choosing your next business.
☑ Yes, we should reject the null hypothesis. Further analysis is needed to choose the best business.
□ No, we should not reject the null hypothesis. You can be indifferent when choosing your next business.
□ No, we should not reject the null hypothesis. You should choose to start a pizza business.
We reject the null, meaning at least two industries have unequal means. Determining which is the lowest requires further analysis.

Problem 2

A climatologist investigates the effect of Solar Radiation on Ozone levels. She takes n = 153 daily observations from May 1st until September 30th. The multiple regression output is:

Coefficient	Estimate	SE	t	Pr(>t)
(Intercept)	-64.34	A	-2.791	B
Solar.R	C	0.023	2.58	0.013

For B, we consider a range of values given the following t-critical points: $t_{0.01} = 2.609$, $t_{0.05} = 1.976$, $t_{0.1} = 1.655$.
Specifically, the three ranges of interest are: • p-value<0.01 • 0.01 <= p-value < 0.05 • 0.05 <= p-value < 0.1 • p-value >= 0.1

Question 22 What is the value for A in the regression output?
□ 0.043
□ 0.92
□ 4.80
☑ 23.05
This is Estimate / t for the '(Intercept)' row

Question 23 What is the value for C in the regression output?
☐ 0.0014
☑ 0.059
☐ 0.39
☐ 112.17
This is SE * t for the 'Solar.R' row.

Question 24 Out of the four ranges mentioned in the question, which range does the p-value for B fall within
☑ p-value<0.01
☐ 0.01<=p-value<0.05
☐ 0.05<=p-value<0.1
☐ p-value>=0.1

|-2.791| > 2.609, so the p-value is less than 0.01

R Data Analysis

For the questions(1-11), you will need to use R and the following dataset:
The dataset was collected from Airbnb with data on listings in the city of Asheville, NC. Here is the data provided for each listing:

- room id: A unique number identifying an Airbnb listing.
- host id: A unique number identifying an Airbnb host.
- room type: One of 'Entire home/apt', 'Private room', or 'Shared room'
- reviews: The number of reviews that a listing has received.
- overall satisfaction: The average rating (out of five) that the listing has received from those visitors who left a review.
- accommodates: The number of guests a listing can accommodate.
- bedrooms: The number of bedrooms a listing offers.
- price: The price (in USD) for a night stay. In early surveys, there may be some values that were recorded by month.

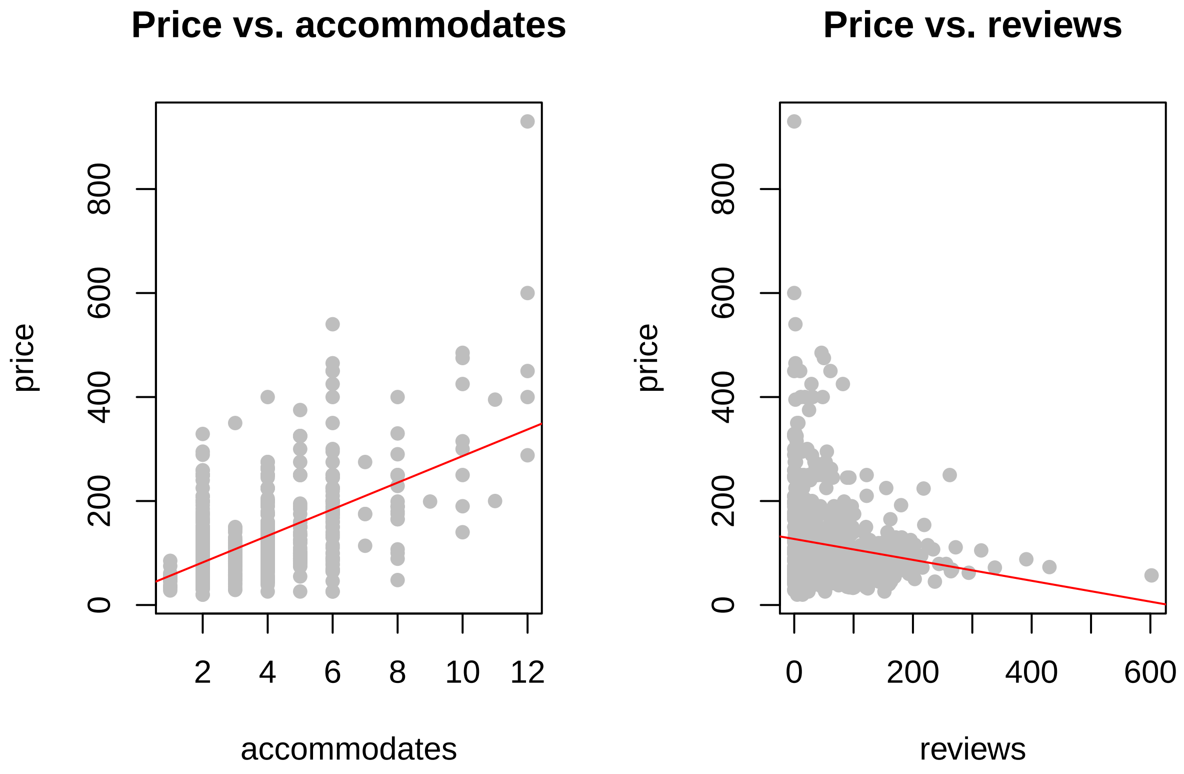
Fit a multiple linear regression model named 'model1' using price as the response variable and the following predicting variables: room type, reviews, overall satisfaction, accommodates, and bedrooms.

```
# Read in the data
house = read.csv("tomslee_airbnb_asheville_1498_2017-07-20.csv", head = TRUE, sep = ",")
# Show the first few rows of data
head(house, 3)
```

```
##    room_id survey_id   host_id   room_type      city reviews
## 1 15771735      1498 101992409 Shared room Asheville       0
## 2 18284194      1498 126414164 Shared room Asheville      32
## 3 18091012      1498 122380971 Shared room Asheville       4
##   overall_satisfaction accommodates bedrooms price
## 1                  0.0            4        1    67
## 2                  5.0            4        1    76
## 3                  4.5            2        1    45
```

Question 1 Create plots of the response, *price*, against two quantitative predictors *accommodates*,and *reviews* . Describe the general trend (direction and form) of each plot

```
# Grid the plots
par(mfrow=c(1,2))
# Plot price vs accommodates
plot(price~accommodates, data=house, main="Price vs. accommodates", col="grey", pch = 16)
abline(lm(price~accommodates, data=house), col="red")
# Plot price vs reviews
plot(price~reviews, data=house, main="Price vs. reviews",col="grey", pch = 16)
abline(lm(price~reviews, data=house), col="red")
```



Response to Question 1: General trend: There appears to be a positive and linear relationship between the response, price, and the predictor, accommodates. There appears to be a slight negative and linear relationship between the response, price, and the predictor, reviews. But we can also observe that there are lots of noise in these two scatters plots, so more analysis would need to be done to determine the strength of the relationships.

Question 2 What is the value of the correlation coefficient for each of the above pair of response and predictor variables? What does it tell you about your comments in Question 1?

```
# Print the correlation coefficients between the predictors and the response
cat("cor(price, accommodates):", cor(house$price, house$accommodates)[1], end="\n")
```

```
## cor(price, accommodates): 0.5886389
```

```
cat("cor(price, reviews):", cor(house$price, house$reviews)[1], end="\n")
```

```
## cor(price, reviews): -0.1532973
```

Response to Question 2: The correlation coefficient between price and accommodates (0.5886389) is the highest of the two groups. This isn't particularly high, but it does communicate that a moderate positive linear relationship between the two variables. The correlation coefficient between price and reviews (-0.1532973) shows a very slight negative linear relationship. These results reinforces that our comments about the general trend for the price vs. accommodates and price vs. reviews plots were correct.

Question 3 Use the *accommodates* as the predictor to build a simple linear regression model for predicting the *price*, named model1. What is the coefficient of *accommodates* in this model?

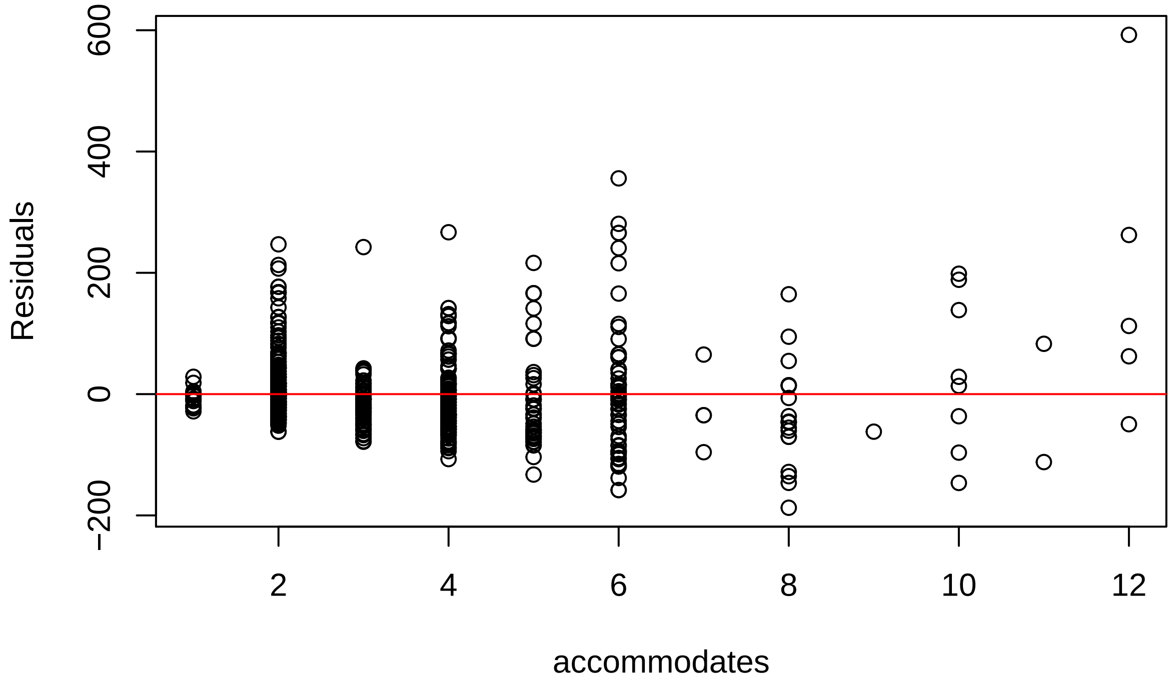
```
model1=lm(price~accommodates, data = house);
summary(model1);
```

```
##
## Call:
## lm(formula = price ~ accommodates, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -187.36  -34.80   -8.12   18.00  592.40
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    30.876      4.632   6.666 4.74e-11 ***
## accommodates   25.560      1.205  21.217  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.82 on 849 degrees of freedom
## Multiple R-squared:  0.3465, Adjusted R-squared:  0.3457
## F-statistic: 450.1 on 1 and 849 DF,  p-value: < 2.2e-16
```

Response to Question 3: The coefficient is 25.560.

Question 4: Assess whether the model assumptions hold, comment on whether there are any apparent departures from the assumptions of the linear regression model. Make sure that you state the model assumptions and assess each one. Each graph may be used to assess one or more model assumptions.

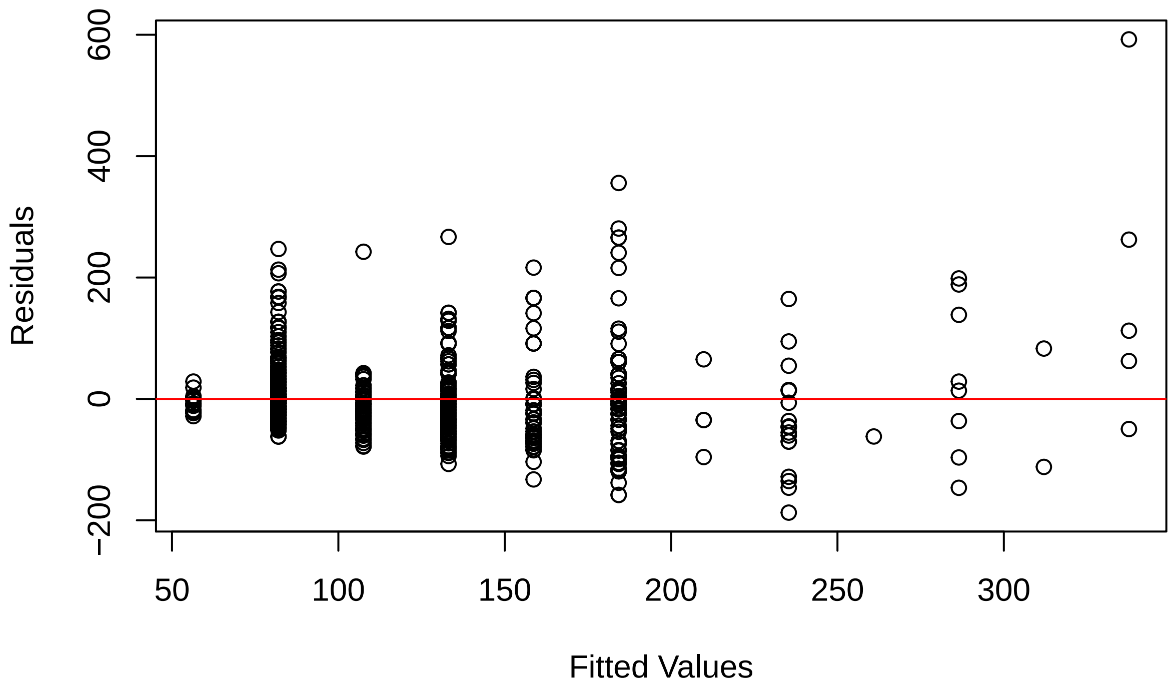
```
## Residuals Vs Predictor Plots
library(MASS)
resids= residuals(model1)
plot(house[,8],resids,xlab="accommodates",ylab="Residuals")
abline(0,0,col="red")
```



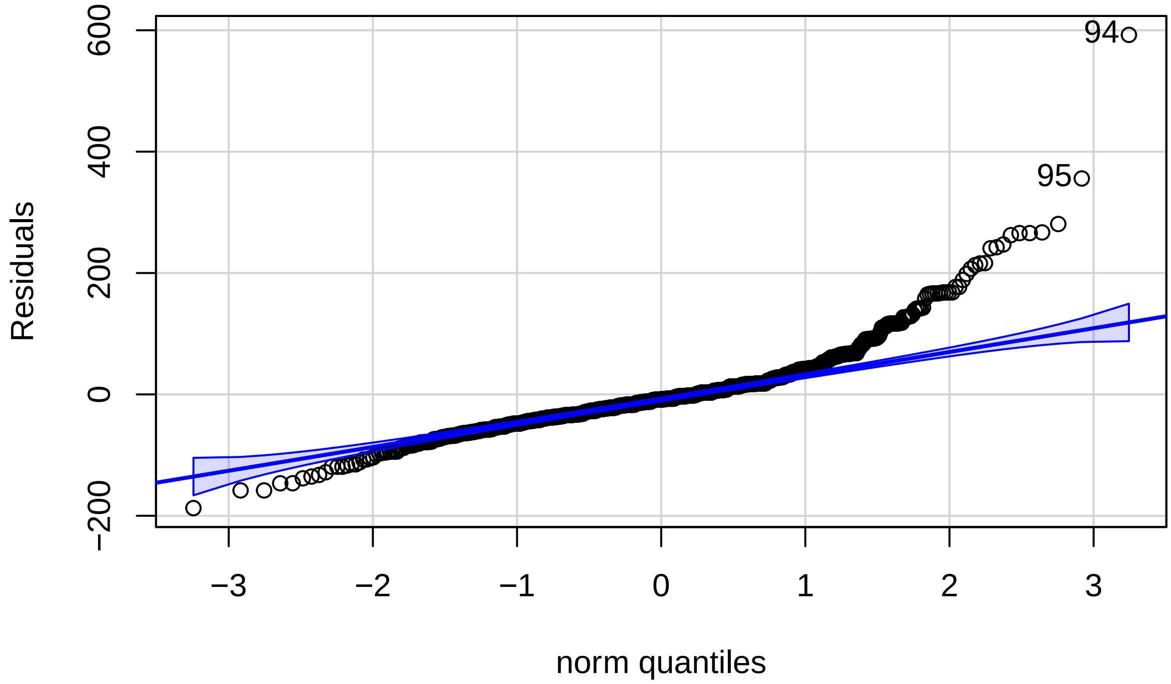
```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.0.5
```

```r
# Residuals Vs Fitted and Q-Q plot
plot(predict(model1), resids, xlab="Fitted Values",ylab="Residuals")
abline(0,0,col="red")
```



```
qqPlot(resids, ylab="Residuals", main = "")
```

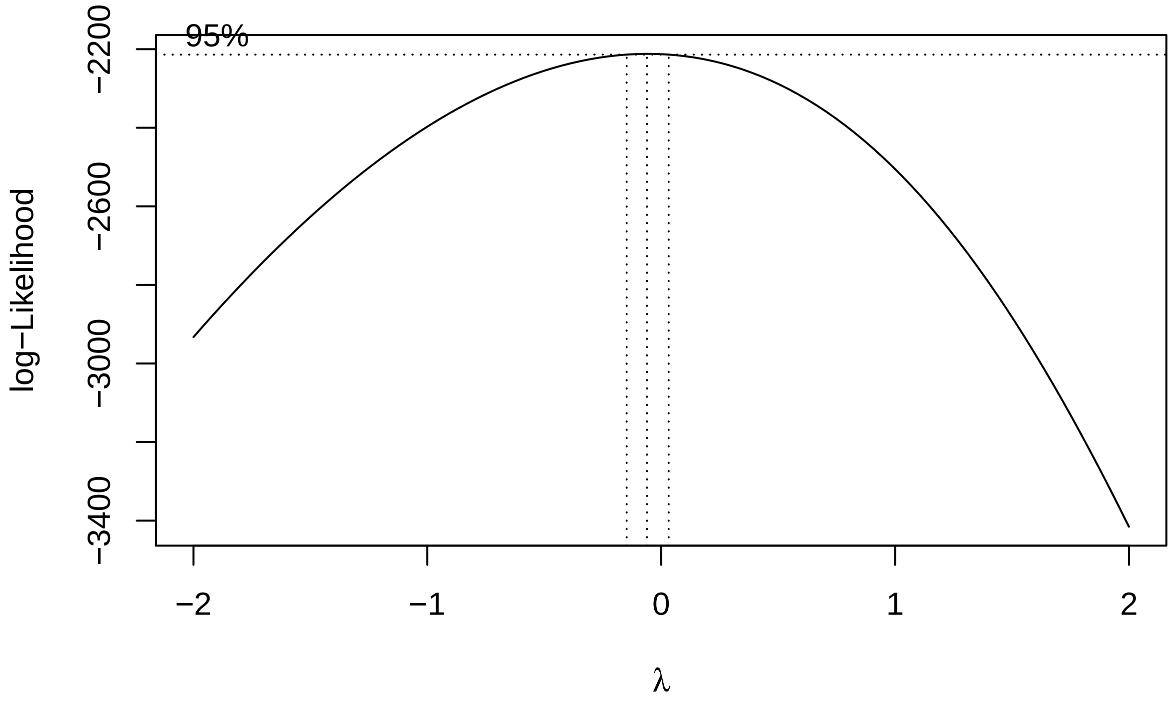


```
## [1] 94 95
```

Response to Question 4: From the residuals/predictor plot, the linearity/mean zero assumption appears to hold reasonably well. Data appears to be symmetrical about the zero line.
From the residuals/fitted plot, the constant variance assumption does not appear to hold. Lower values have smaller variance than higher values.
From the residuals/fitted plot, the uncorrelated error assumption holds, there are no apparent clusters of residuals.
From the qq plot, the normality assumption does not appear to hold. The data appears to be skewed to the right.

Question 5 For improving the fitness, we can use a box-cox transformation. Find the optimal lambda value rounded to the nearest half integer. Report this best lambda and corresponding transformation.

```
bc = boxcox(model1);
```



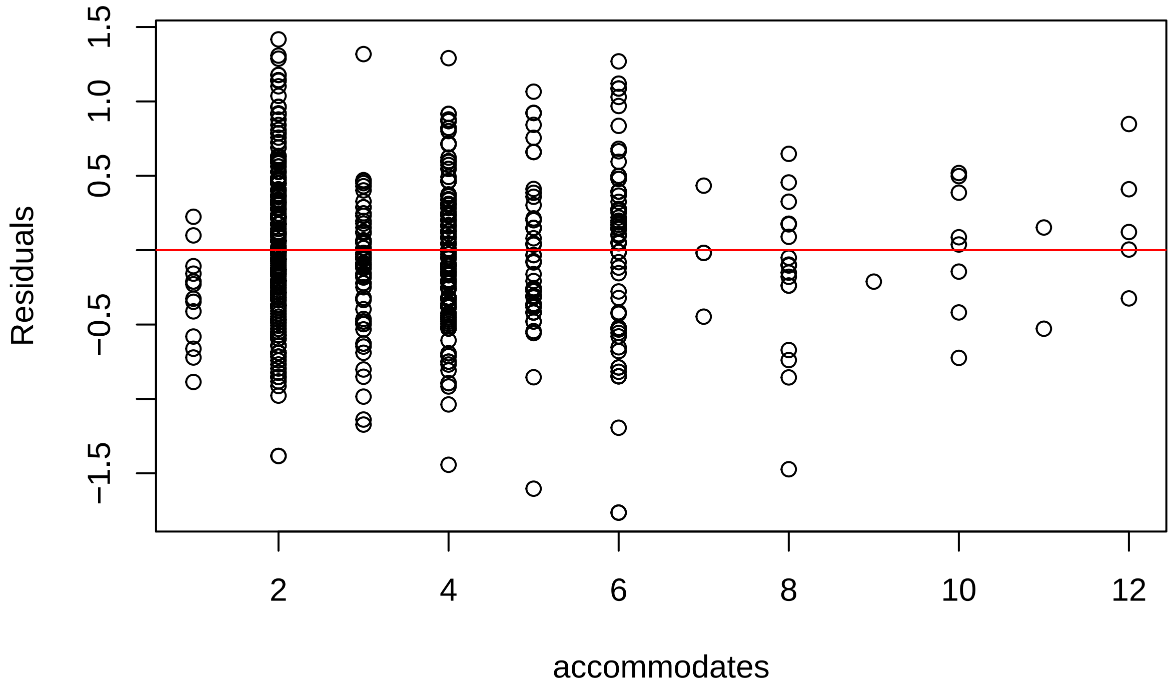
```
opt.lambda = bc$x[which.max(bc$y)]
cat("Optimal lambda:", round(opt.lambda/0.5)*0.5, end="\n")
```

```
## Optimal lambda: 0
```

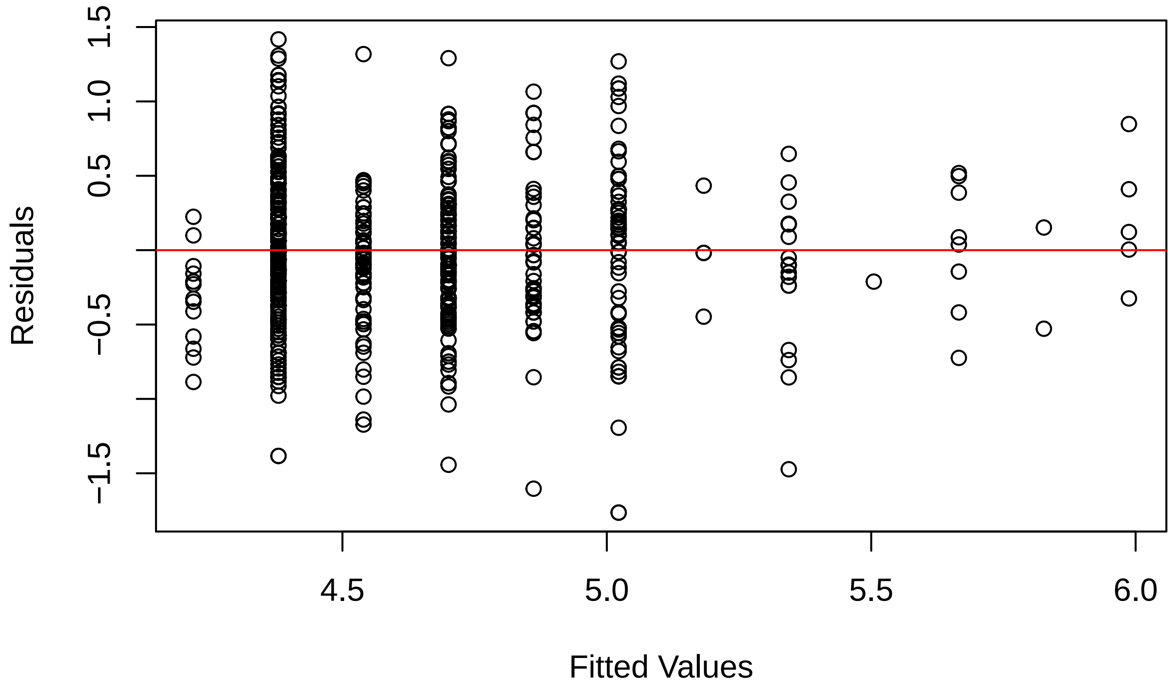
Response to Question 5 The optimal value of lambda should be 0. The optimal lambda value is zero, suggesting that the log of the response may improve constant variance and the normality.

Question 6 Use this optimal lambda value to transform the response variable. Build a new simple linear regression model, named model2, with the transformed response and the predictor *accommodates*. Similar to Question 4, check the model assumptions. Does this model seem a better fit?

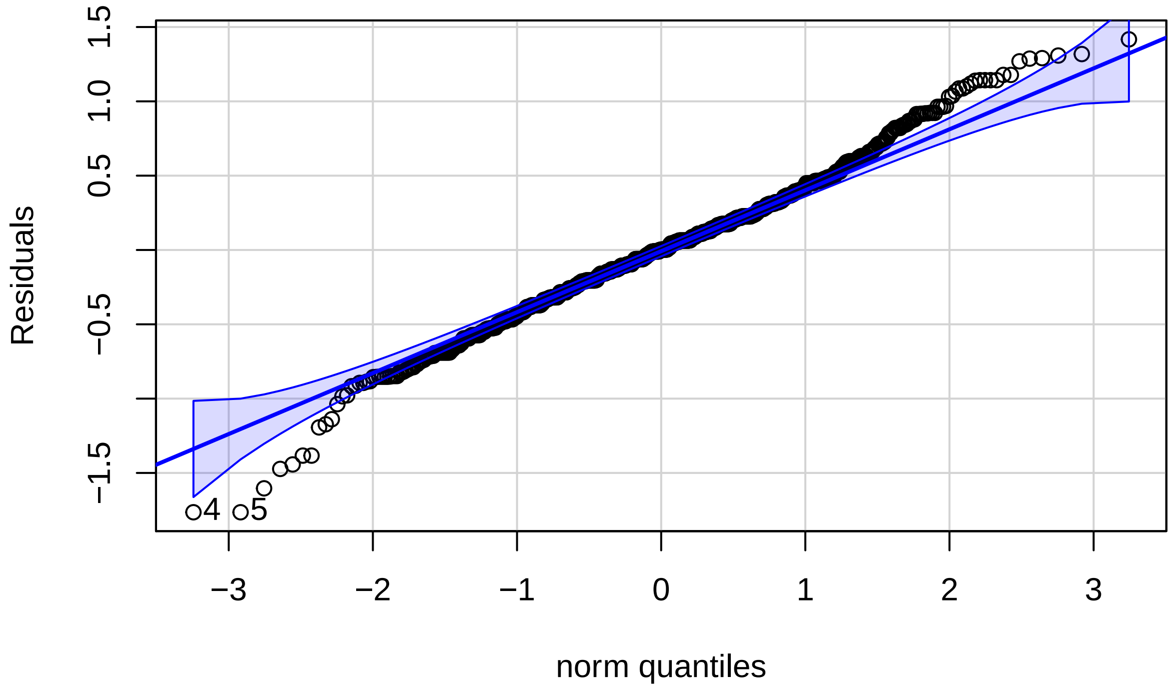
```
model2 = lm(log(price)~accommodates, data = house);
resids= residuals(model2)
plot(house[,8],resids,xlab="accommodates",ylab="Residuals")
abline(0,0,col="red")
```



```
plot(predict(model2), resids, xlab="Fitted Values",ylab="Residuals")
abline(0,0,col="red")
```



```
qqPlot(resids, ylab="Residuals", main = "")
```



```
## [1] 4 5
```

Response to Question 6 From the residuals/predictor plot, the linearity/mean zero assumption appears to hold reasonably well. Data appears to be symmetrical about the zero line.
From the residuals/fitted plot, the constant variance assumption appear to hold. Lower values and higher values have the same variance.
From the residuals/fitted plot, the uncorrelated error assumption holds, there are no apparent clusters of residuals.
From the qq plot, the normality assumption appear to hold.
So we can conclude that, after the transformation, the model assumptions can be better met.

Question 7 Fit a multiple linear regression model, using price as the response variable and the following predicting variables: room type, reviews, overall satisfaction, accommodates, and bedrooms. Which coefficients (including intercept) are statistically significant at the 99% confidence level?

```
# Build the model
model3 = lm(price ~ room_type + reviews + overall_satisfaction + accommodates + bedrooms, data = house)
# Show the 99% confidence intervals
confint(model3,level = 0.99)
```

```
##                                0.5 %       99.5 %
## (Intercept)               57.2982887  99.19648106
## room_typePrivate room    -45.6753949 -20.88675705
## room_typeShared room    -149.2845561 -37.38566358
## reviews                  -0.1599122   0.02653578
## overall_satisfaction     -8.7732808  -2.25692842
## accommodates              7.1044650  16.97542961
## bedrooms                 17.6616970  40.27329765
```

Response to Question 7:
Coefficient | 0.5 % | 99.5 % |
————|———|——|
(Intercept) | 57.2982887| 99.19648106
room_typePrivate room | -45.6753949 |-20.88675705
room_typeShared room | -149.2845561 | -37.38566358
reviews | -0.1599122 | 0.02653578 |
overall_satisfaction | -8.7732808 | -2.25692842
accommodates | 7.1044650 | 16.97542961
bedrooms | 17.6616970 | 40.27329765
☑ Intercept
☑ room type of Private room
☑ room type of Shared room
□ reviews
☑ overall satisfaction
☑ accomodates
☑ bedrooms

Question 8 What is the estimated coefficient for room type = "Private Room" in this MLR model?

```r
model3$coefficients['room_typePrivate room']
```

```
## room_typePrivate room
##             -33.28108
```

Response to Question 8: room_typePrivate room -33.28108

Question 9 What is the interpretation for the estimated coefficient for room type = "Private Room"?
Response to Question 9 A listing for a private room has an estimated cost of 33.28 USD less than an entire home/apt, holding all other variables constant.

Question 10 Report the coefficient of determination for your MLR model and give a concise interpretation of this value.

```r
# Extract R^2
cat("R^2:",summary(model3)$r.squared)
```

```
## R^2: 0.4353298
```

Response to Question 10: R^2 is 0.4353298 or 43.53%. We can interpret this as 43.53% of the variation in the response is explained by the predictors in the model.

Question 11 Using your MLR model, make a prediction for a listing on Airbnb in Asheville with the following factors:
bedrooms = 1, accommodates = 2, reviews = 92, overall_satisfaction = 3.5, and room_type= 'Private

room'.
What is your predicted price for such a listing and the corresponding 95% prediction interval?

```
new_data = data.frame(bedrooms=1, accommodates=2, reviews=92,
overall_satisfaction = 3.5, room_type= 'Private room')
predict(model3, new_data, interval="prediction", level=0.95)
```

```
##        fit       lwr      upr
## 1 72.57552 -46.28007 191.4311
```

Response to Question 11:
fit | lwr | upr |
—-|—–|—–|
72.57552 | -46.28007 | 191.4311