ASSIGNMENT COVER SHEET

Research School of Finance, Actuarial Studies and Statistics
ANU College of Business and Economics
Australian National University
Canberra ACT 0200 Australia
www.anu.edu.au

This cover sheet should be attached to the front of your assignment, which must be submitted online using Turnitin.

Student ID For group assignments, list each student's ID	U7446204		
Course Code	STAT6038		
Course Name	Regression Modelling		
Assignment number	Assignment-1		
Assignment Topic	Analysing Used Car Prices with Simple Linear Regression		
Lecturer	Dr Insha Ullah		
Tutor			
Tutorial (day and time)	Monday 10:00AM		
Word count	3157	Due Date	31 Mar 2023 - 15:00
Date Submitted	30 Mar 2023	Extension Granted	

I declare that this work:

☑ upholds the principles of academic integrity, as defined in the ANU Policy: Code of Practice for Student Academic Integrity;

☑ is original, except where collaboration (for example, group work) has been authorised in writing by the course convener in the course outline and/or Wattle site;

☑ is produced for the purposes of this assessment task and has not been submitted for assessment in any other context, except where authorised in writing by the course convener;

☑ gives appropriate acknowledgement of the ideas, scholarship and intellectual property of others insofar as these have been used;

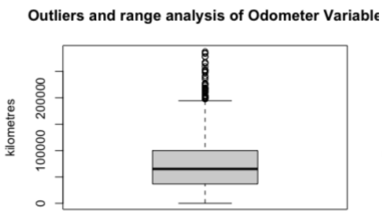
☑ in no part involves copying, cheating, collusion, fabrication, plagiarism or recycling.

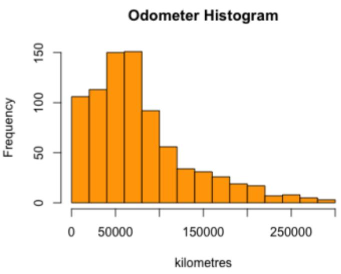
Initials

For group assignments, each student must initial (digital is acceptable provided all participants have confirmed)

Assignment-u7446204: Analyzing Used Car Prices with Simple Linear Regression

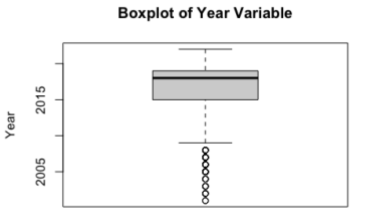
a) Exploring the predictor variable- Odometer

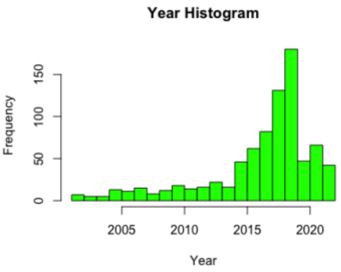




According to boxplot, the outlier in Odometer's data, these data are greater than Q3+1.5IQR. In addition to that, mean(77490.0709)> median(65138), we roughly judge that the data is right skewed at this time, And, the range of Odometer is (8 , 287552).

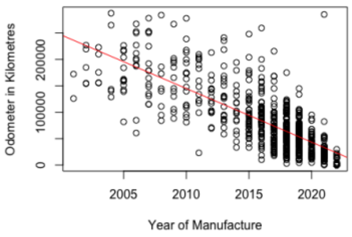
```
Omean <- mean(carsales$Odometer)
Omedian<-median(carsales$Odometer)
```





```
range(carsales$Year)
Ymean<- mean(carsales$Year)
Ymedian<- median(carsales$Year)
```

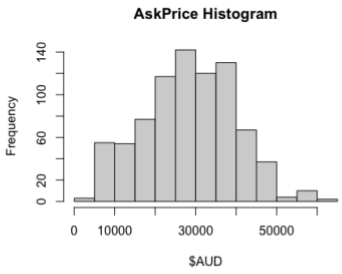
Based on the above information, mean(2016.6198) < median(2018), we roughly judge that the data is left skewed at this time , the range of Year is (2001 , 2022)

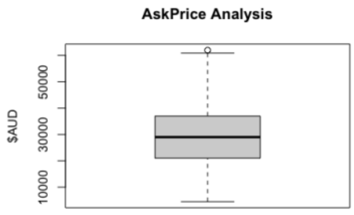


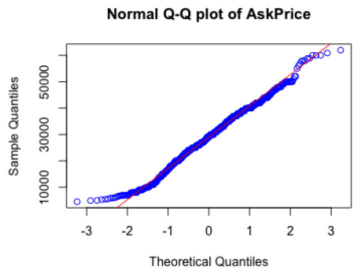
The scatter plot shows a negative correlation between the two variables, with a correlation coefficient of -0.7660442

```
r_YO <- cor(carsales$Year, carsales$Odometer)     ## [1] -0.7660442
```

b) Appropriate graphic diagnostics and descriptive statistics for AskPrice







According to the box plot, it is believed that there is an outlier.

After observation, it is possible to say the data of AskPrice may be normally distributed, and then we should perform the Shapiro-

Wilk test on the data and compare the p-value(=3.327e-05) with the significant level(given alpha = 5%)

```
shapiro.test(carsales$AskPrice)
```

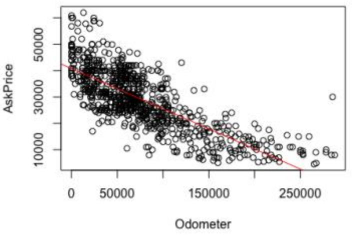
While the p-vlue = 3.327e-05 which is far less than the alpha- 0.05, so we should say the data of AskPrice is not normal distribution. Meanwhile, we should use the Q-Q plot to observe whether the variable AskPrice is normal distribution or not

Since AskPrice has only 818 data, even though in the Shapiro-Wilk test, we believe that the data does not conform to a normal distribution based on the comparison of p-value with alpha, we can still roughly assume that AskPrice data conforms to a normal distribution based on Q-Q plots, etc.

c) Analysis to assess the correlation between the two variables, AskPrice and Odometer

We aim to analyze the reasonableness of car offers and therefore consider offers as outcomes, depending on which factors they are related to. Askprice is more of a dependent variable than Odometer.

According to the analysis, the two show a negative correlation. The smaller the Odometer value, the less the car is used and the less the mileage, the less the car is worn, the less the depreciation, and therefore the higher the value of the car itself, the higher the Ask Price, in line with the reality of the transaction. According to cor. test, it is considered that t-vlaue is -34.616 p-value is 2.2e-16 < 0.05. Therefore, it is considered that there is a significant correlation between the two variables

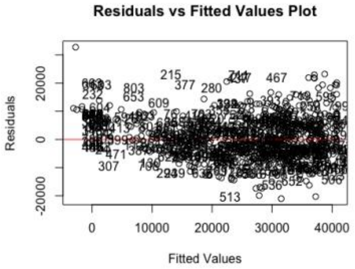


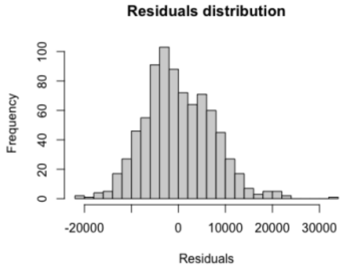
d)

According to the case information, AskPrice is the response variable and Odometer is the predictor variable.

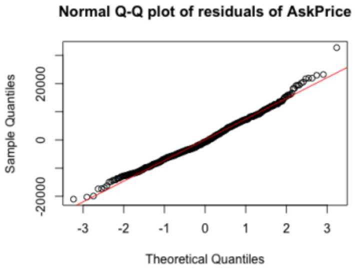
We create a plot of the residuals against the fitted values: the graphs show that the residuals are not randomly distributed around 0, indicating that the linearity and variance chi-squared of the model are not well satisfied.

The horizontal axis in this graph is the y-value (Fitted value) and the vertical axis is the residuals. If the residuals tend to increase or decrease as the y-value increases, or if the distribution of the residuals more closely resembles a quadratic curve, then it means that the original data may not be linear. At this point, we can do some transformations such as logarithm, exponential, square root, etc., and then perform linear regression.

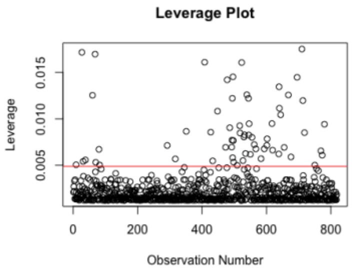


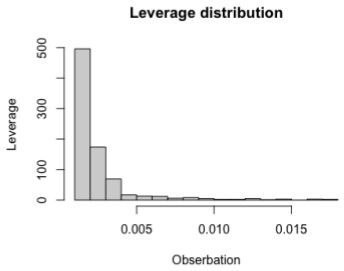


As to the Q-Q plot of the residuals: Based on the Q-Q plot and histogram, it can be initially concluded that the residuals have the characteristics of normal distribution.



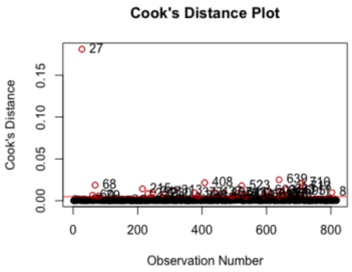
We also need to consider the leverage plot: This plot is used to detect the leverage value of each data point. If a data point has a large leverage value, it indicates that the data point is very far from the other data points in the X-space and may have a large impact on the estimation of the regression coefficients. Therefore, it can be considered to exclude these data points when fitting the model to avoid the effect of these data points on the model fit.





Based on the graphs, it is clear that there are many data points in the model_OA that have large leverage values and are far from other data points that may have a large impact on the regression coefficients, or that the regression model may not be very appropriate using SLR.

As to creating a bar plot of Cook's distances for each observation:



If the data points have a small impact on the model, their Cook's distance is small. If the Cook's distance of some data points is large, it indicates that these data points have a large influence on the model and are called outliers. Therefore, these data points can be considered for exclusion when fitting the model.

In the figure, the outlier is large at the point 27.

e)

```
Fit a simple linear regression model
model_OA <- lm(carsales$AskPrice~carsales$Odometer, data = carsales)
anova(model_OA)

## Analysis of Variance Table
##
## Response: carsales$AskPrice
##                    Df     Sum Sq    Mean Sq F value    Pr(>F)
## carsales$Odometer   1 6.2975e+10 6.2975e+10  1198.3 < 2.2e-16 ***
## Residuals         816 4.2885e+10 5.2555e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With the anova table as above, we can do the F-test.

F test value is used to test whether correlation between the AskPrice and Odometer is zero or not

H0: beta1 =0 and Ha: beta1 !=0, f_test_OA=MSR_OA/MSE_OA

F value is 1198.262(close to the value 1198.3), which is far from the 1 and we should reject the H0, so the beta1 != 0 ,and there is a liner relationship between the AskPrice(y) and Odometer(x), and the model between two variables are significant.

We also can calculate the coefficient of derminatiion R^2:

```
r2_OA<- SSR_OA/SSTO_OA ## [1] 0.5948889
```

Finally, we can use the summary function to summarize and verify:

```
summary(model_OA)

##
## Call:
## lm(formula = carsales$AskPrice ~ carsales$Odometer, data = carsales)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21016  -4900   -735   5015  32738
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.077e+04  4.252e+02   95.87   <2e-16 ***
## carsales$Odometer -1.525e-01  4.406e-03  -34.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7249 on 816 degrees of freedom
## Multiple R-squared:  0.5949, Adjusted R-squared:  0.5944
## F-statistic:  1198 on 1 and 816 DF,  p-value: < 2.2e-16
```

f)

When we suppose the liner relationship between AskPrice and Odometer, based on the result of part b), we could say the intercept of the liner model is beta0=4.077e+04, and the standard error is sigma=4.252e+02, which means when the Odometer is zero and the AskPrice is 4.077e+04. We can use the t-test to to determine if beta0 differ significantly from zero.

H0: beta0=0 , Ha: beta0 != 0 and the significant level is 5%

beta0_t = estimated beta0/ standard error{beta0}=4.077e+04/4.252e+02

t-test should be two-side test and the degree of freedom id n-2 (= 818-2=816)

```
beta0_ttest<- qt(1-0.05/2, (length(carsales$Odometer)-2))
beta0_t > beta0_ttest
```

According to the TEST, beta0_t value is 95.88 while the beta0_ttest value is 1.96 when the significant level is 5%, t value is larger than the t test value (beta0_t > beta0_ttest), and we should reject H0, and the p-value is 2e-16 which is less than then 0.05, which means beta0 is significant different from zero.

According to the result of part b), we could say the estimated beta1=-1.525e-01, that is the slop of the liner model. And the standard error is 4.406e-03. We can use the t-test to to determine if beta1 differ significantly from zero.

H0: beta1=0 , Ha: beta1 != 0 and the significant level is 5%

beta1_t = estimated beta1/ standard error{beta1}=1.525e-01/4.406e-03

```
beta1_t <--1.525e-01/4.406e-03
beta1_ttest<- qt(1-0.05/2,(length(carsales$Odometer)-2) )
abs(beta1_t) >beta1_ttest
```

According to the TEST,   beta1_t value is -34.61 while the beta1_ttest value is 1.96 when the significant level is 5%, the absolute value of beta1_t is larger than the t test value (beta1_t > beta1_ttest), and we should reject H0, and the p-value is 2e-16 which is less than then 0.05, which means beta1 is significant different from zero.

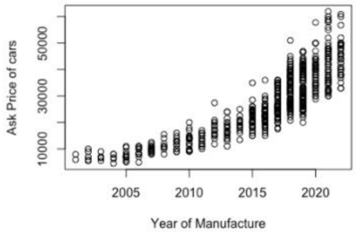
g)

We create a plot of AskPrice and Year, which show a positive correlation.

After that , we calculate the correlation value of two variable:

```
cor(carsales$Year, carsales$AskPrice)    [1] 0.8440862
```

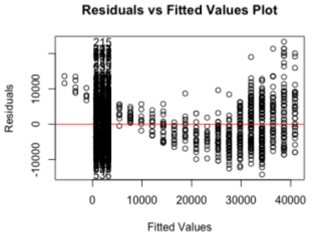
The correlation estimated value is 0.8440862. According to the chart, we expect that the two should be positively correlated.

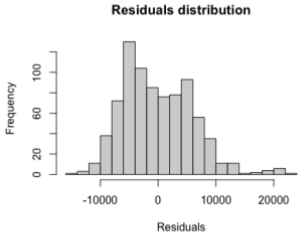


h)

1) Create a plot of the residuals against the fitted values

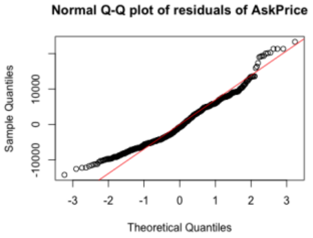
   The graphs show that the residuals are not randomly distributed around 0, indicating that the linearity and variance chi-squared of the model are not well satisfied





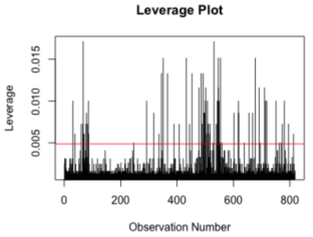
2) Create a normal Q-Q plot of the residuals

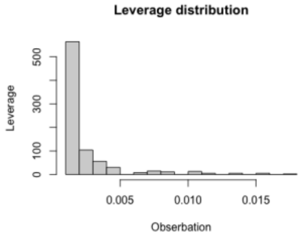
   We often use Q-Q plots to check whether the residuals conform to a normal distribution. Obviously there are deviations between the points and the line, thus indicating that the model is able to capture some features of the data or that there is an unorthodox error.



3) Create a bar plot of the leverages for each observation

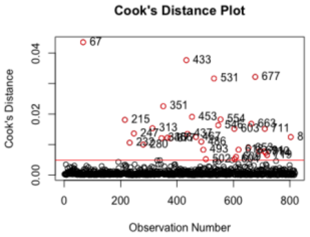
   Based on the graphs, it is clear that there are many data points in the model_YA that have large leverage values and are far from other data points that may have a large impact on the regression coefficients, or that the regression model may not be very appropriate using SLR.





4) Create a bar plot of Cook's distances for each observation

   The graph is used to detect the impact of each data point on the model. If the data points have a small impact on the model, their Cook's distance is small. If some data points have a large Cook's distance, they have a large influence on the model and may be outliers or outliers. Therefore, these data points can be considered for exclusion when fitting the model. In the figure, the outliers are 67, 433, 531, and 677.



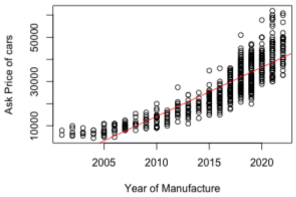
i)

```
model_YA<-lm(carsales$AskPrice~ carsales$Year)
anova(model_YA)
## Analysis of Variance Table
```

```
##
## Response: carsales$AskPrice
##               Df     Sum Sq    Mean Sq F value    Pr(>F)    
## carsales$Year   1 7.5423e+10 7.5423e+10  2022.1 < 2.2e-16 ***
## Residuals     816 3.0437e+10 3.7300e+07                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Therefore we will perform a test for F-value and we test that we expect the two to show a positive correlation, beta1>0. And the correlation estimated value is 0.8440862.

```
cor(carsales$Year, carsales$AskPrice)
```



So we assume that: H0: beta1=0 , Ha: beta1 !=0   while the F-value = MSR/MSE. According to the result of the anova of model_YA, the F-value is 2022.1 which is far from 1. Therefore, we should reject H0 and the beta1 != 0 and the model_YA is significant.

```
r2_YA <-(cor(carsales$AskPrice, carsales$Year))^2
```

we use the r2_YA to present the value of R^2= 0.71248 , which means 71.248% variation of Askprice can be explained by the model.

Finally, we can use the summary function to summarize and verify:

```
summary(model_YA)
## Call:
## lm(formula = carsales$AskPrice ~ carsales$Year)
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -14270.3  -4698.0   -797.4   4653.3  23319.4 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   -4.438e+06  9.934e+04  -44.68   <2e-16 ***
## carsales$Year  2.215e+03  4.926e+01   44.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 6107 on 816 degrees of freedom
## Multiple R-squared:  0.7125, Adjusted R-squared:  0.7121 
## F-statistic:  2022 on 1 and 816 DF,  p-value: < 2.2e-16
```

j)

When we suppose the liner relationship between AskPrice and Year, based on the result above, we could say the intercept of the liner model is beta0_YA=-4.438e+06, and the standard error is sigma=9.934e+04, which means when the Year is zero and the AskPrice is -4.438e+06. We can use the t-test to to determine if beta0 differ significantly from zero.

H0: beta0=0 , Ha: beta0 != 0 and the significant level is 5%

beta0_t = estimated beta0/ standard error{beta0}=4.438e+06/9.934e+04

t-test should be two-side test and the degree of freedom id n-2 (= 818-2=816)

```
beta0_YA_ttest<- qt(1-0.05/2, (length(carsales$Year)-2))
abs(beta0_YA_t )> beta0_YA_ttest
```

According to the TEST,   beta0_YA_t value is -44.67 while the beta0_YA_ttest value is 1.96 when the significant level is 5%, the absolute value of beta0_YA_t is larger than the t test value (beta0_YA_t > beta0_YA_ttest), and we should reject H0, and the p-value is 2e-16 which is less than then 0.05, which means beta0 is significant different from zero.

According to the result above, we could say the estimated beta1_YA=2.215e+03, that is the slop of the liner mofel. And the stnadard error is 4.926e+01. We can use the t-test to to determine if beta1 differ significantly from zero.

H0: beta1=0 , Ha: beta1 != 0 and the significant level is 5%

beta1_t = estimated beta1/ standard error{beta1}=2.215e+03/4.926e+01

```
beta1_YA_ttest<- qt(1-0.05/2,(length(carsales$Year)-2) )
abs(beta1_YA_t) >beta1_YA_ttest
```

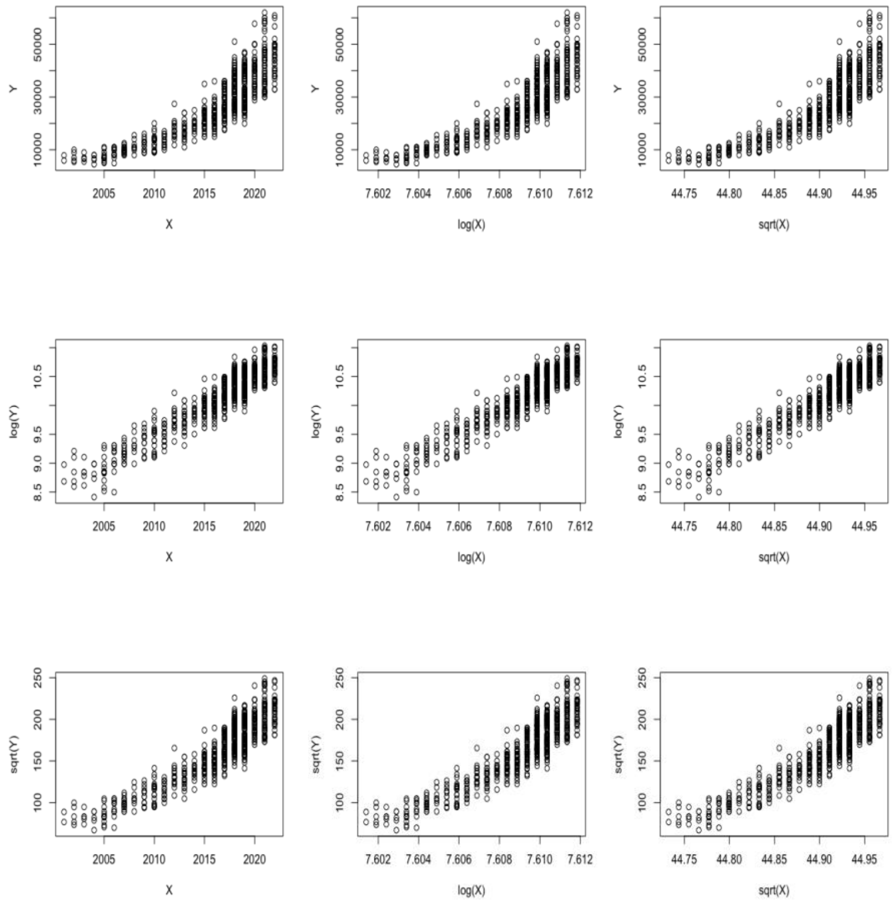
According to the TEST, beta1_YA_t value is 44.97 while the beta1_YA_ttest value is 1.96 when the significant level is 5%, the absolute value of beta1_YA_t is larger than the t test value (beta1_YA_t >beta1_YA_ttest), and we should reject H0, and the p-value is 2e-16 which is less than then 0.05, which means beta1 is significant different from zero.

k)

```
X <- carsales$Year
Y <- carsales$AskPrice
```

It is not really appropriate to use SLR model between X,Y, so we can choose to use SLR model separately for AskPrice and Year doing exponential or logarithmic processing.

```
plot(X,Y),plot(log(X),Y) ,plot(sqrt(X),Y),plot(X,log(Y)),plot(log(X),log(Y)),plot(sqrt(X),log(Y),plot(X,sqrt(Y)),plot(log(X),
sqrt(Y)),plot(sqrt(X),sqrt(Y))
```



According to the comparison of graphs and correlation coefficients, when log(Y), X basically has a more stable linear relationship with log(Y), regardless of the raw data or the processed form. Therefore, it is considered more appropriate to use the SLR model to represent between X and log(Y).

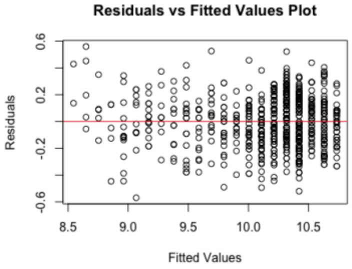
And, as the year of the independent variable X, the change should indeed be a simple increase or decrease of one unit, AskPrice can instead have multiple forms of change.

```
model_XY<- lm(log(Y)~X)
summary(model_XY)
```

```
## Call:
## lm(formula = log(Y) ~ X)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56968 -0.12119 -0.00205  0.14235  0.55952
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -199.98491    3.00498  -66.55   <2e-16 ***
## X              0.10421    0.00149   69.94   <2e-16 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.1847 on 816 degrees of freedom
## Multiple R-squared:  0.857,  Adjusted R-squared:  0.8568
## F-statistic:  4891 on 1 and 816 DF,  p-value: < 2.2e-16
```

According to the result of the summary of model_XY, when the Year increase 1 year, the log(AskPrice) will increase 0.10421.

l)



According to the question stem, we should evaluate the selling price of the car. We would recommend the sample car with a smaller value of residual, and vice versa we would refuse to recommend the sample car with a positive and larger value of residual. The Question need us to condider the parts (d), (h), and (k)

```
carsales[which.max(resid(model_OA)),] carsales[which.min(resid(model_OA)),]
```

IN part(b) , we should choose to recommend the car 79, which AskPrice is $AUD 10500 with 60648km while we should reject recommending the car 27 , which AskPrice is $AUD 30000 with 285255km.

```
carsales[which.max(resid(model_YA)),] carsales[which.min(resid(model_YA)),]
```

IN part(h) , we should choose to recommend the car 536, which AskPrice is $AUD 19950 with the year of manufacture is 2019 while we should reject recommending the car 215 , which AskPrice is $AUD 61970 with the year of manufacture is 2021

```
carsales[which.max(resid(model_XY)),] carsales[which.min(resid(model_XY)),]
```

IN part(k) , we should choose to recommend the car 496, which AskPrice is $AUD 4900 with the year of manufacture is 2006 while we should reject recommending the car 433 , which AskPrice is $AUD 9990 with the year of manufacture is 2002.

m)

A confidence interval is a range that gives the probability that the true value may lie within that range, and a confidence level of 95% is usually used.It can be explained that using a given linear regression model and the year and mileage of the vehicle as the new data, we can create a confidence interval that will tell us where the sales price of the car will fall at a 95% confidence level. The sales price will fall in (9.991056 , 10.01813)

```
confidence_interval_XY<-predict(model_XY,newdata= data.frame(X=2015), interval="confidence", level = 0.95)
```

```
exp(confidence_interval_XY)

##        fit      lwr      upr
## 1 22127.83 21830.35 22429.37
```

Usually, if a linear regression model uses a logarithmic transformation, the predictions will also be in logarithmic form, so that if we want to obtain confidence intervals for the actual values , we would choose to use the exp()function for gain the data. According to the data of exp(predict()), at a 95% confidence level the sales price falls between (21830.35 , 22429.37)

```
pred_interval_XY <- predict(model_XY, newdata =data.frame(X=2015), interval = "prediction", level = 0.95)
exp(pred_interval_XY)

##        fit      lwr      upr
## 1 22127.83 15393.71 31807.86
```

The prediction interval is a range that gives a prediction for the new data, taking into account the effects of model error and random error. The prediction interval is (9.641714 , 10.36747), then we can say that the car will be sold in this range with a 95% confidence level.

Similarly, we have logged the data in model_XY, so we will choose the exp() function to reprocess the data. Thus we get, based on the existing data prediction, we think that the predicted eligible Askprice should fall in (15393.71 , 31807.86) with a 95% confidence level.

Recommendation:

If it costs us less to acquire the car than the minimum price $ANU 15393.71,, then we will recommend less than the minimum price if the seller finds the profit margin acceptable.

Also, the prices within the price range are the normal range of market prices $AUD(15393.71, 31807.86)and the car seller should choose the price range that fits the forecast based on the size of the acquisition cost and the profit margin.

If the price is higher than the maximum price $ANUD 31807.86, the car seller's offer will not be competitive in the market and we will not recommend it.

For log-transformed models, the model output is a forecast of the logarithmic price. Therefore, when giving price recommendations, you need to reverse transform the prediction to the original price. Using predict(interval="prediction") will provide you with a prediction interval containing the actual prices of the new observations. Use the exp() function to convert log prices to raw prices, so you should use exp(predict(interval="prediction")) to calculate the suggested price interval. This interval will contain a 95% probability that the actual price will fall within this interval.

I hope to give suggestions for price ranges based on the model AskPrice~ Year+Odometer:

```
model <- lm(AskPrice ~ Year + Odometer)
conf_interval <- predict(model, newdata =data.frame(Year = 2015, Odometer = 100000), interval = "confidence", level = 0.95)
pred_interval <- predict(model, newdata = data.frame(Year = 2015, Odometer = 100000), interval = "prediction", level = 0.95
)
```

Using predict(interval = "prediction") will give an interval that has a 95% probability of containing the actual price of a new observation. Therefore, this interval will help you determine the recommended price range. In this case, the recommended price range should be the lower and upper limits of the resulting prediction interval.

The prediction interval is (13808.25 , 36190.15), then we can say that the car will be sold in this range with a 95% confidence level.

Based on the 'AskPrice~ Year + Odometer' model, I would recommend the purchase of an eligible car when its price is below $ANU 13808.25, and I would not recommend it when it is above $ANU 36190.15.

Appendix:

```
carsales <- read.csv("carsales.csv", header = T)
attach(carsales)
head(carsales)

##   Year   Model AskPrice Odometer Transmission  Power
## 1 2021 Corolla    35990    28929    Automatic Hybrid
## 2 2022 Corolla    37888    14020    Automatic Hybrid
## 3 2013   Camry    23950   128870    Automatic Hybrid
## 4 2022 Corolla    42499     2400    Automatic Hybrid
## 5 2015   Camry    22975    99143    Automatic Petrol
## 6 2022   Camry    49950        9    Automatic Hybrid
```

##Analysing Used Car Prices with Simple Linear Regression

##a)Exploring the predictor variable- Odometer

```
Omean <- mean(carsales$Odometer)
Omedian<-median(carsales$Odometer)
Osd<- sd(carsales$Odometer)
Omin<- min(carsales$Odometer)
Omax<- max(carsales$Odometer)
range(carsales$Odometer)

## [1]      8 287552

Omax-Omin

## [1] 287544

IQR(carsales$Odometer, na.rm = TRUE)

## [1] 63292.5

# Boxplot-graphic method:
boxplot(carsales$Odometer, type = "box" , main = "Outliers and range analysis of Odometer Variable" , ylab= "kilometres")

# Hhistogram plot
hist(carsales$Odometer, breaks = 20, col = "orange", main = "Odometer Histogram",xlab = "kilometres")

# Based on the above information, mean > median,we roughly judge that the data is right skewed at this time


# Year vriable
Ysd<- sd(carsales$Year)
Ymin<- min(carsales$Year)
Ymax<- max(carsales$Year)
range(carsales$Year)

## [1] 2001 2022

Ymean<- mean(carsales$Year)
Ymedian<- median(carsales$Year)
# Boxplot-graphic method:
boxplot(carsales$Year, type = "box" , main = "Boxplot of Year Variable" ,  ylab= "Year")
```

```
# Hhistogram plot
hist(carsales$Year, breaks = 20, col = "green", main = "Year Histogram",xlab = "Year")

# Based on the above information, mean < median,we roughly judge that the data is left skewed at this time
r_YO <- cor(carsales$Year, carsales$Odometer)
r_YO

## [1] -0.7660442

abs(r_YO)

## [1] 0.7660442

plot(carsales$Odometer~carsales$Year, xlab = "Year of Manufacture", ylab = "Odometer in Kilometres")
abline(lm(carsales$Odometer~carsales$Year), col="red")

# In fact, both Year  and Odometer are very much like observed values, so it is difficult to say cause and effect. It is far-fetched to say that Odometer is more objective, so in the plot, it is more preferable to consider Year as the independent variable and Odometer as the dependent variable
# At this point, there MAY be a negative linear correlation between Year and Odometer.
# Testing Hypothsese:
# H0:beta1 =0, VS
# H1:beta1 < 0, alpha=0.05,
length(carsales$Year)

## [1] 818

length(carsales$Odometer)

## [1] 818

df.correlation<-length(carsales$Year)-2
t.correlation<-abs(r_YO)*sqrt(df.correlation/(1-r_YO^2))
t.correlation

## [1] 34.04326

t.correlation.test<-qt(1-0.05,df.correlation)
t.correlation.test

## [1] 1.646723

#Based on the significance level of 5%, and the degrees of freedom df.correlation=816,t.correlation=34.04326 >t.correlation.test= 1.646723, the null hypothesis H0 is REJECTED, indicating that there is a significant negative linear correlation between Year and Odometer, and the correlation between the year of manufacture and odometer reading in kilometers is less than zero.
```

##b)

```
summary(carsales$AskPrice)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4500   20996   28990   28948   36990   61970

hist(carsales$AskPrice, breaks = 20, main = "AskPrice Histogram",xlab = "$AUD")
```

```
boxplot(carsales$AskPrice, type = "box" , main = "AskPrice Analysis" ,  ylab= "$AUD")

# According to the box plot, it is believed that there is an outlier.
# After observation, it is possible to say the data of AskPrice may be normally distributed, and then we should perform the Shapiro-Wilk test on the data and compare the p-value with the significant level(given alpha = 5%)
shapiro.test(carsales$AskPrice)

## 
##  Shapiro-Wilk normality test
## 
## data:  carsales$AskPrice
## W = 0.99034, p-value = 3.327e-05

#While the p-vlue = 3.327e-05 which is far less than the alpha- 0.05, so we should say the data of AskPrice is not normal distrobution
# If use the Q-Q plot
qqnorm(carsales$AskPrice, main = "Normal Q-Q plot of AskPrice", col = "blue")
qqline(carsales$AskPrice, col= "red")

length(carsales$AskPrice)

## [1] 818

# Since AskPrice has only 818 data, even though in the Shapiro-Wilk test, we believe that the data does not conform to a normal distribution based on the comparison of p-value with alpha, we can still roughly assume that AskPrice data conforms to a  normal distribution based on Q-Q plots, etc.
```

##c) # Analysis to assess the correlation between the two variables, AskPrice and Odometer

```
cor(carsales$Odometer, carsales$AskPrice)

## [1] -0.7712904

model_OA <- lm(carsales$AskPrice~carsales$Odometer)
plot(carsales$Odometer, carsales$AskPrice, xlab ="Odometer" , ylab ="AskPrice")
abline(model_OA, col="red")

cor.test(carsales$Odometer, carsales$AskPrice)

## 
##  Pearson's product-moment correlation
## 
## data:  carsales$Odometer and carsales$AskPrice
## t = -34.616, df = 816, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7976651 -0.7419712
## sample estimates:
##        cor 
## -0.7712904

summary(model_OA)
```

```
##
## Call:
## lm(formula = carsales$AskPrice ~ carsales$Odometer)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21016  -4900   -735   5015  32738
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.077e+04  4.252e+02   95.87   <2e-16 ***
## carsales$Odometer -1.525e-01  4.406e-03  -34.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7249 on 816 degrees of freedom
## Multiple R-squared:  0.5949, Adjusted R-squared:  0.5944
## F-statistic:  1198 on 1 and 816 DF,  p-value: < 2.2e-16

# We aim to analyze the reasonableness of car offers and therefore consider offers as outcomes, depending on which factors they are related to. Askprice is more of a dependent variable than Odometer.
# According to the analysis, the two show a negative correlation. The smaller the Odometer value, the less the car is used and the less the mileage, the less the car is worn, the less the depreciation, and therefore the higher the value of the car itself, the higher the Ask Price, in line with the reality of the transaction
# According to cor. test, it is considered that t-vlaue is -34.616 p-value is 2.2e-16 < 0.05. Therefore, it is considered that there is a significant correlation between the two variables
```

##d) # According to the case information, AskPrice is the response variable and Odometer is the predictor variable.

```
# Fit a simple linear regression model
model_OA <- lm(carsales$AskPrice~carsales$Odometer, data = carsales)

# Create a plot of the residuals against the fitted values
dif_points_OA <- which(abs(resid(model_OA)) > 2)
plot(model_OA$fitted.values, resid(model_OA), main = "Residuals vs Fitted Values Plot", xlab = "Fitted Values", ylab = "Residuals")
abline(h = mean(resid(model_OA)) , col = "red")
text(x = carsales$Odometer[dif_points_OA], y = resid(model_OA)[dif_points_OA], labels = dif_points_OA, )

# The graphs show that the residuals are not randomly distributed around 0, indicating that the linearity and variance chi-squared of the model are not well satisfied.
# The horizontal axis in this graph is the y-value (Fitted value) and the vertical axis is the residuals. If the residuals tend to increase or decrease as the y-value increases, or if the distribution of the residuals more closely resembles a quadratic curve, then it means that the original data may not be linear. At this point, we can do some transformations such as logarithm, exponential, square root, etc., and then perform linear regression.

# Create a normal Q-Q plot of the residuals
qqnorm(resid(model_OA), main = "Normal Q-Q plot of residuals of AskPrice")
qqline(resid(model_OA), col = "red")

hist(resid(model_OA), breaks = 20, main = "Residuals distribution",xlab = "Residuals", ylab = "Frequency")
```

```
#Based on the Q-Q plot and histogram, it can be initially concluded that the residuals have the characteristics of normal distribution.

# Create a bar plot of the leverages for each observation
plot(hatvalues(model_OA), main = "Leverage Plot", xlab = "Observation Number", ylab = "Leverage")
abline( h = 2*mean(hatvalues(model_OA)) , col ="red")

hist(hatvalues(model_OA), breaks = 20, main = "Leverage distribution",xlab = "Obserbation", ylab = "Leverage")

# This plot is used to detect the leverage value of each data point. If a data point has a large leverage value, it indicates that the data point is very far from the other data points in the X-space and may have a large impact on the estimation of the regression coefficients. Therefore, it can be considered to exclude these data points when fitting the model to avoid the effect of these data points on the model fit.
# Based on the graphs, it is clear that there are many data points in the model_OA that have large leverage values and are far from other data points that may have a large impact on the regression coefficients, or that the regression model may not be very appropriate using SLR.

# Create a bar plot of Cook's distances for each observation
plot(cooks.distance(model_OA), main = "Cook's Distance Plot", xlab = "Observation Number", ylab = "Cook's Distance")
abline(h = 4/length(cooks.distance(model_OA)), col= "red")
cooksd_OA<-cooks.distance(model_OA)
cooks_abline <- 4/length(cooksd_OA)
cooks_high <- which(cooksd_OA > cooks_abline)
points(cooks_high, cooksd_OA[cooks_high], pch = 1, col = "red")
text(cooks_high, cooksd_OA[cooks_high], cooks_high, pos = 4)

# If the data points have a small impact on the model, their Cook's distance is small. If the Cook's distance of some data points is large, it indicates that these data points have a large influence on the model and are called outliers. Therefore, these data points can be considered for exclusion when fitting the model.
# In the figure, the outlier is large at the point 27.
summary(model_OA)

##
## Call:
## lm(formula = carsales$AskPrice ~ carsales$Odometer, data = carsales)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21016  -4900   -735   5015  32738
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.077e+04  4.252e+02   95.87   <2e-16 ***
## carsales$Odometer -1.525e-01  4.406e-03  -34.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7249 on 816 degrees of freedom
## Multiple R-squared:  0.5949, Adjusted R-squared:  0.5944
## F-statistic:  1198 on 1 and 816 DF,  p-value: < 2.2e-16
```

```
#In fact, we can perform exponential or logarithmic transformation on x and y, and check if there is a linear relationship between them based on the graphs.
# Logarithm of y-AskPrice
model_OA1 <- lm(log(carsales$AskPrice) ~ carsales$Odometer)
plot(model_OA1)
```

```
summary(model_OA1)
```

```
##
## Call:
## lm(formula = log(carsales$AskPrice) ~ carsales$Odometer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14252 -0.14806  0.01783  0.19126  1.56514
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.071e+01  1.676e-02  638.97   <2e-16 ***
## carsales$Odometer -6.881e-06  1.736e-07  -39.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2857 on 816 degrees of freedom
## Multiple R-squared:  0.6581, Adjusted R-squared:  0.6577
## F-statistic:  1571 on 1 and 816 DF,  p-value: < 2.2e-16
```

```
# Take the Logarithm of x-Odometer
model_OA2 <- lm(carsales$AskPrice ~ log(carsales$Odometer))
plot(model_OA2)
```

```
summary(model_OA2)
```

```
##
## Call:
## lm(formula = carsales$AskPrice ~ log(carsales$Odometer))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25636.9  -5724.5   -170.8   6938.4  27687.2
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)               81202       2332   34.82   <2e-16 ***
## log(carsales$Odometer)    -4840        214  -22.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8931 on 816 degrees of freedom
## Multiple R-squared:  0.3852, Adjusted R-squared:  0.3845
## F-statistic: 511.3 on 1 and 816 DF,  p-value: < 2.2e-16
```

```
# Square the y-AskPrice transform
model_OA3 <- lm(sqrt(carsales$AskPrice) ~ carsales$Odometer)
plot(model_OA3)

summary(model_OA3)

##
## Call:
## lm(formula = sqrt(carsales$AskPrice) ~ carsales$Odometer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.529 -13.572  -0.842  15.340 110.618
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.050e+02  1.262e+00  162.45   <2e-16 ***
## carsales$Odometer -4.993e-04  1.308e-05  -38.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.52 on 816 degrees of freedom
## Multiple R-squared:  0.6411, Adjusted R-squared:  0.6407
## F-statistic:  1458 on 1 and 816 DF,  p-value: < 2.2e-16

# However, according to the correlation graphs of the following model1-3, it is difficult to find a suitable transformation
of x and y that satisfies a simple linear relationship.
```

##e)

```
s<- 7249
df_sse<-816
SSE_OA<- s^2*df_sse
SSE_OA

## [1] 42879168816

cor(carsales$AskPrice, carsales$Odometer)

## [1] -0.7712904

SSTO_OA<- SSE_OA/(1-(cor(carsales$AskPrice, carsales$Odometer)^2))
SSR_OA<- SSTO_OA-SSE_OA
MSR_OA<- SSR_OA/1
MSE_OA<- SSE_OA/df_sse
#The coefficient of derminatiion- R^2
r2_OA<- SSR_OA/SSTO_OA
r2_OA

## [1] 0.5948889

r2_OA == (cor(carsales$AskPrice, carsales$Odometer))^2

## [1] TRUE
```

```
summary(model_OA)

##
## Call:
## lm(formula = carsales$AskPrice ~ carsales$Odometer, data = carsales)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21016  -4900   -735   5015  32738
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.077e+04  4.252e+02   95.87   <2e-16 ***
## carsales$Odometer -1.525e-01  4.406e-03  -34.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7249 on 816 degrees of freedom
## Multiple R-squared:  0.5949, Adjusted R-squared:  0.5944
## F-statistic:  1198 on 1 and 816 DF,  p-value: < 2.2e-16

#F test value is used to test whether correlation between the AskPrice and Odometer is zero or not
# H0: beta1 =0 and Ha: beta1 !=0
f_test_OA<- MSR_OA/MSE_OA
f_test_OA

## [1] 1198.262

# F value is 1198.262(close to the value 1198.3 ), which is far from the 1 and we should reject the H0, so the beta1 != 0 ,and there is a liner relationship between the AskPrice(y) and Odometer(x), and the model between two variables are significant.
anova(model_OA)

## Analysis of Variance Table
##
## Response: carsales$AskPrice
##                    Df     Sum Sq    Mean Sq F value    Pr(>F)
## carsales$Odometer   1 6.2975e+10 6.2975e+10  1198.3 < 2.2e-16 ***
## Residuals         816 4.2885e+10 5.2555e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

##f)

```
# while we suppose the liner relationship between AskPrice and Odometer, based on the result of part b), we could say the intercept of the liner model is beta0=4.077e+04, and the standard error is sigma=4.252e+02, which means when the Odometer is zero and the AskPrice is 4.077e+04. We can use the t-test to to determine if beta0 differ significantly from zero
# H0: beta0=0 , Ha: beta0 != 0 and the significant level is 5%
# beta0_t = estimated beta0/ standard error{beta0}
beta0_t<- 4.077e+04/4.252e+02
# t-test should be two-side test and the degree of freedom id n-2 (= 818-2=816)
```

```
beta0_ttest<- qt(1-0.05/2, (length(carsales$Odometer)-2))
beta0_t > beta0_ttest

## [1] TRUE

# According to the TEST,  beta0_t value is 95.88 while the beta0_ttest value is 1.96 when the significant level is 5%, t value is larger than the t test value (beta0_t > beta0_ttest), and we should reject H0, and the p-value is 2e-16 which is less than then 0.05, which means beta0 is significant different from zero.


#According to the result of part b), we could say the estimated beta1=-1.525e-01, that is the slop of the liner mofel. And the stnadard error is 4.406e-03. We can use the t-test to to determine if beta1 differ significantly from zero.
# H0: beta1=0 , Ha: beta1 != 0 and the significant level is 5%
# beta1_t = estimated beta1/ standard error{beta1}
beta1_t <--1.525e-01/4.406e-03
beta1_ttest<- qt(1-0.05/2,(length(carsales$Odometer)-2) )
abs(beta1_t) >beta1_ttest

## [1] TRUE

# According to the TEST,  beta1_t value is -34.61 while the beta1_ttest value is 1.96 when the significant level is 5%, the absolute value of beta1_t is larger than the t test value (beta1_t > beta1_ttest), and we should reject H0, and the p-value is 2e-16 which is less than then 0.05, which means beta1 is significant different from zero.
```

##g)

```
plot(carsales$AskPrice~carsales$Year, xlab = "Year of Manufacture", ylab = "Ask Price of cars")

cor(carsales$Year, carsales$AskPrice)

## [1] 0.8440862

# the correlation estimated value is 0.8440862
# According to the chart, we expect that the two should be positively correlated.
```

##h)

```
model_YA<-lm(carsales$AskPrice~ carsales$Year)
# Create a plot of the residuals against the fitted values
dif_points_YA <- which(abs(resid(model_YA)) > 2)
plot(model_YA$fitted.values, resid(model_YA), main = "Residuals vs Fitted Values Plot", xlab = "Fitted Values", ylab = "Residuals")
abline(h = mean(resid(model_YA)) , col = "red")
text(x = carsales$Year[dif_points_YA], y = resid(model_YA)[dif_points_YA], labels = dif_points_YA )

#The graphs show that the residuals are not randomly distributed around 0, indicating that the linearity and variance chi-squared of the model are not well satisfied


# Create a normal Q-Q plot of the residuals
qqnorm(resid(model_YA), main = "Normal Q-Q plot of residuals of AskPrice")
qqline(resid(model_YA), col = "red")

hist(resid(model_YA), breaks = 20, main = "Residuals distribution",xlab = "Residuals", ylab = "Frequency")
```

```
# We often use Q-Q plots to check whether the residuals conform to a normal distribution. Obviously there are deviations between the points and the line, thus indicating that the model is able to capture some features of the data or that there is an unorthodox error.

# Create a bar plot of the leverages for each observation
plot(hatvalues(model_YA), type = "h", main = "Leverage Plot", xlab = "Observation Number", ylab = "Leverage")
abline( h = 2*mean(hatvalues(model_YA)) , col ="red")

hist(hatvalues(model_YA), breaks = 20, main = "Leverage distribution",xlab = "Obserbation", ylab = "Leverage")

# Based on the graphs, it is clear that there are many data points in the model_YA that have large leverage values and are far from other data points that may have a large impact on the regression coefficients, or that the regression model may not be very appropriate using SLR.

# Create a bar plot of Cook's distances for each observation
plot(cooks.distance(model_YA), main = "Cook's Distance Plot", xlab = "Observation Number", ylab = "Cook's Distance")
abline(h = 4/length(cooks.distance(model_YA)), col= "red")
cooksd_YA<-cooks.distance(model_YA)
cooks_abline_YA <- 4/length(cooksd_YA)
cooks_high_YA <- which(cooksd_YA > cooks_abline_YA)
points(cooks_high_YA, cooksd_YA[cooks_high_YA], pch = 1, col = "red")
text(cooks_high_YA, cooksd_YA[cooks_high_YA], cooks_high_YA, pos = 4)

# The graph is used to detect the impact of each data point on the model. If the data points have a small impact on the model, their Cook's distance is small. If some data points have a large Cook's distance, they have a large influence on the model and may be outliers or outliers. Therefore, these data points can be considered for exclusion when fitting the model.
# In the figure, the outliers are 67, 433, 531, and 677.
```

##i)

```
summary(model_YA)

##
## Call:
## lm(formula = carsales$AskPrice ~ carsales$Year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14270.3  -4698.0   -797.4   4653.3  23319.4
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.438e+06  9.934e+04  -44.68   <2e-16 ***
## carsales$Year  2.215e+03  4.926e+01   44.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6107 on 816 degrees of freedom
## Multiple R-squared:  0.7125, Adjusted R-squared:  0.7121
## F-statistic:  2022 on 1 and 816 DF,  p-value: < 2.2e-16

anova(model_YA)
```

```
## Analysis of Variance Table
##
## Response: carsales$AskPrice
##                Df     Sum Sq    Mean Sq F value    Pr(>F)
## carsales$Year   1 7.5423e+10 7.5423e+10  2022.1 < 2.2e-16 ***
## Residuals     816 3.0437e+10 3.7300e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Therefore we will perform a test for F-value and we test that we expect the two to show a positive correlation, beta1>0.
cor(carsales$Year, carsales$AskPrice)
```

```
## [1] 0.8440862
```

```
# the correlation estimated value is 0.8440862
plot(carsales$AskPrice~carsales$Year, xlab = "Year of Manufacture", ylab = "Ask Price of cars")
abline(lm(carsales$AskPrice~carsales$Year), col="red")
```

```
# So we assume that:
# H0: beta1=0 , Ha: beta1 !=0  while the F-value = MSR/MSE. According to the result of the anova of model_YA, the F-value is 2022.1 which is far from 1. Therefore, we should reject H0 and the beta1 != 0 and the model_YA is significant.
r2_YA <-(cor(carsales$AskPrice, carsales$Year))^2
# we use the r2_YA to present the value of R^2= 0.71248 , which means 71.248% variation of Askprice can be explained by the model.
```

##j)

```
# while we suppose the liner relationship between AskPrice and Year, based on the result above, we could say the intercept of the liner model is beta0_YA=-4.438e+06, and the standard error is sigma=9.934e+04, which means when the Year is zero and the AskPrice is -4.438e+06. We can use the t-test to to determine if beta0 differ significantly from zero
# H0: beta0=0 , Ha: beta0 != 0 and the significant level is 5%
# beta0_t = estimated beta0/ standard error{beta0}
beta0_YA_t<- -4.438e+06/9.934e+04
# t-test should be two-side test and the degree of freedom id n-2 (= 818-2=816)
beta0_YA_ttest<- qt(1-0.05/2, (length(carsales$Year)-2))
abs(beta0_YA_t )> beta0_YA_ttest
```

```
## [1] TRUE
```

```
# According to the TEST,  beta0_YA_t value is -44.67 while the beta0_YA_ttest value is 1.96 when the significant level is 5%, the absolute value of beta0_YA_t is larger than the t test value (beta0_YA_t > beta0_YA_ttest), and we should reject H0, and the p-value is 2e-16 which is less than then 0.05, which means beta0 is significant different from zero.

#According to the result above, we could say the estimated beta1_YA=2.215e+03, that is the slop of the liner mofel. And the stnadard error is 4.926e+01. We can use the t-test to to determine if beta1 differ significantly from zero.
# H0: beta1=0 , Ha: beta1 != 0 and the significant level is 5%
# beta1_t = estimated beta1/ standard error{beta1}
beta1_YA_t <-2.215e+03/4.926e+01
beta1_YA_ttest<- qt(1-0.05/2,(length(carsales$Year)-2) )
abs(beta1_YA_t) >beta1_YA_ttest
```

```
## [1] TRUE
```

```
# According to the TEST,  beta1_YA_t value is 44.97 while the beta1_YA_ttest value is 1.96 when the significant level is 5%,
 the absolute value of beta1_YA_t is larger than the t test value (beta1_YA_t >beta1_YA_ttest), and we should reject H0, and
 the p-value is 2e-16 which is less than then 0.05, which means beta1 is significant different from zero.
```

##k)

```
X <- carsales$Year
Y <- carsales$AskPrice
#It is not really appropriate to use SLR model between X,Y, so we can choose to use SLR model separately for AskPrice and Ye
ar doing exponential or logarithmic processing.
plot(X,Y)

plot(log(X),Y)

plot(sqrt(X),Y)

plot(X,log(Y))

plot(log(X),log(Y))

plot(sqrt(X),log(Y))

plot(X,sqrt(Y))

plot(log(X),sqrt(Y))

plot(sqrt(X),sqrt(Y))

# while we should consider the correlation between X and Y
cor(X,Y)

## [1] 0.8440862

cor(log(X),Y)

## [1] 0.8437685

cor(sqrt(X),Y)

## [1] 0.8439275

cor(X,log(Y))

## [1] 0.9257546

cor(log(X),log(Y))

## [1] 0.9257181

cor(sqrt(X),log(Y))

## [1] 0.9257366

cor(X,sqrt(Y))

## [1] 0.8937971

cor(log(X),sqrt(Y))
```

```
## [1] 0.8936067

cor(sqrt(X),sqrt(Y))

## [1] 0.8937021

# According to the comparison of graphs and correlation coefficients, when log(Y), X basically has a more stable linear relationship with log(Y), regardless of the raw data or the processed form. Therefore, it is considered more appropriate to use the SLR model to represent between X and log(Y).
# And, as the year of the independent variable X, the change should indeed be a simple increase or decrease of one unit, Ask Price can instead have multiple forms of change.
model_XY<- lm(log(Y)~X)
summary(model_XY)

##
## Call:
## lm(formula = log(Y) ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56968 -0.12119 -0.00205  0.14235  0.55952
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -199.98491    3.00498  -66.55   <2e-16 ***
## X              0.10421    0.00149   69.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1847 on 816 degrees of freedom
## Multiple R-squared:  0.857,  Adjusted R-squared:  0.8568
## F-statistic:  4891 on 1 and 816 DF,  p-value: < 2.2e-16

# According to the result of the summary of model_XY, when the Year increase 1 year, the log(AskPrice) will increase 0.10421.
```

##l)

```
residuals_XY <- resid(model_XY)
# Create the residual plot of part k) based on the log(Y) and X
dif_points_XY <- which(abs(resid(model_XY)) > 2)
plot(model_XY$fitted.values, residuals_XY, main = "Residuals vs Fitted Values Plot", xlab = "Fitted Values", ylab = "Residuals")
abline(h = mean(residuals_XY), col = "red")

if(length(dif_points_XY) > 0) {
  text(x = X[dif_points_XY], y = residuals_XY[dif_points_XY], labels = dif_points_XY)
}
#According to the question stem, we should evaluate the selling price of the car. We would recommend the sample car with a smaller value of residual, and vice versa we would refuse to recommend the sample car with a positive and larger value of residual.
```

```
# The Question need us to condider the parts (d), (h), and (k)
which.max(resid(model_OA))

## 27
## 27

carsales[which.max(resid(model_OA)),]

##    Year Model AskPrice Odometer Transmission  Power
## 27 2021 Camry    30000   285255    Automatic Petrol

which.min(resid(model_OA))

## 79
## 79

carsales[which.min(resid(model_OA)),]

##    Year   Model AskPrice Odometer Transmission  Power
## 79 2006 Corolla    10500    60648    Automatic Petrol

# IN part(b) , we should choose to recommend the car 79, which AskPrice is $AUD 10500   with 60648km while we should reject recommending the car 27 , which AskPrice is $AUD 30000 with 285255km.
which.max(resid(model_YA))

## 215
## 215

carsales[which.max(resid(model_YA)),]

##     Year Model AskPrice Odometer Transmission  Power
## 215 2021 Camry    61970    13145    Automatic Hybrid

which.min(resid(model_YA))

## 536
## 536

carsales[which.min(resid(model_YA)),]

##     Year   Model AskPrice Odometer Transmission  Power
## 536 2019 Corolla    19950    29985    Automatic Petrol

# IN part(h) , we should choose to recommend the car 536, which AskPrice is $AUD 19950  with the year of manufacture  is 2019 while we should reject recommending the car 215 , which AskPrice is $AUD 61970 with the year of manufacture  is 2021.
which.max(resid(model_XY))

## 433
## 433

carsales[which.max(resid(model_XY)),]

##     Year   Model AskPrice Odometer Transmission  Power
## 433 2002 Corolla     9990   184767    Automatic Petrol

which.min(resid(model_XY))
```

```
## 496
## 496

carsales[which.min(resid(model_XY)),]

##     Year Model AskPrice Odometer Transmission  Power
## 496 2006 Camry     4900   267300    Automatic Petrol

# IN part(k) , we should choose to recommend the car 496, which AskPrice is $AUD 4900   with the year of manufacture  is 2006 while we should reject recommending the car 433 , which AskPrice is $AUD 9990  with the year of manufacture  is 2002.
```

##m)

```
confidence_interval_XY<-predict(model_XY,newdata= data.frame(X=2015), interval="confidence", level = 0.95)
confidence_interval_XY

##        fit      lwr      upr
## 1 10.00459 9.991056 10.01813

#A confidence interval is a range that gives the probability that the true value may lie within that range, and a confidence level of 95% is usually used.It can be explained that using a given linear regression model and the year and mileage of the vehicle as the new data, we can create a confidence interval that will tell us where the sales price of the car will fall at a 95% confidence level.
# the sales price will fall in (9.991056 , 10.01813)
exp(confidence_interval_XY)

##        fit      lwr      upr
## 1 22127.83 21830.35 22429.37

# Usually, if a linear regression model uses a logarithmic transformation, the predictions will also be in logarithmic form, so that if we want to obtain confidence intervals for the actual values , we would choose to use the exp()function for gain the data.
# According to the data of exp(predict()), at a 95% confidence level the sales price falls between (21830.35 , 22429.37)
pred_interval_XY <- predict(model_XY, newdata =data.frame(X=2015), interval = "prediction", level = 0.95)
pred_interval_XY

##        fit      lwr      upr
## 1 10.00459 9.641714 10.36747

#The prediction interval is a range that gives a prediction for the new data, taking into account the effects of model error and random error. The prediction interval is (9.641714 , 10.36747), then we can say that the car will be sold in this range with a 95% confidence level.
exp(pred_interval_XY)

##        fit      lwr      upr
## 1 22127.83 15393.71 31807.86

# Similarly, we have logged the data in model_XY, so we will choose the exp() function to reprocess the data. Thus we get, based on the existing data prediction, we think that the predicted eligible Askprice should fall in (15393.71 , 31807.86) with a 95% confidence level.
# Recommendation: Based on the question, I would choose the expected price od predicted value to determine the lower bound of the car price. Therefore, based on the known data, I would recommend the purchase of an eligible car when its price is below $ANU 15393.71, and I would not recommend it when it is above  $ANU 31807.86.
#For log-transformed models, the model output is a forecast of the logarithmic price. Therefore, when giving price recommen
```

```
dations, you need to reverse transform the prediction to the original price. Using predict(interval="prediction") will provide you with a prediction interval containing the actual prices of the new observations. Use the exp() function to convert log prices to raw prices, so you should use exp(predict(interval="prediction")) to calculate the suggested price interval. This interval will contain a 95% probability that the actual price will fall within this interval.

# I hope to give suggestions for price ranges based on the model AskPrice~ Year+Odometer:
Year <- carsales$Year
Odometer <- carsales$Odometer
AskPrice<-carsales$AskPrice
model <- lm(AskPrice ~ Year + Odometer)
new_data <- data.frame(Year = Year, Odometer = Odometer)
new_row <- data.frame(Year = 2015, Odometer = 100000)
new_data <- rbind(new_data, new_row)
conf_interval <- predict(model, newdata =data.frame(Year = 2015, Odometer = 100000), interval = "confidence", level = 0.95)
conf_interval

##       fit      lwr      upr
## 1 24999.2 24576.93 25421.47

pred_interval <- predict(model, newdata = data.frame(Year = 2015, Odometer = 100000), interval = "prediction", level = 0.95
)
pred_interval

##       fit      lwr      upr
## 1 24999.2 13808.25 36190.15

# Using predict(interval = "prediction") will give an interval that has a 95% probability of containing the actual price of a new observation. Therefore, this interval will help you determine the recommended price range. In this case, the recommended price range should be the lower and upper limits of the resulting prediction interval.
#The prediction interval is (13808.25 , 36190.15), then we can say that the car will be sold in this range with a 95% confidence level.
# Based on the 'AskPrice~ Year + Odometer' model,I would recommend the purchase of an eligible car when its price is below $ANU 13808.25, and I would not recommend it when it is above  $ANU 36190.15.
```