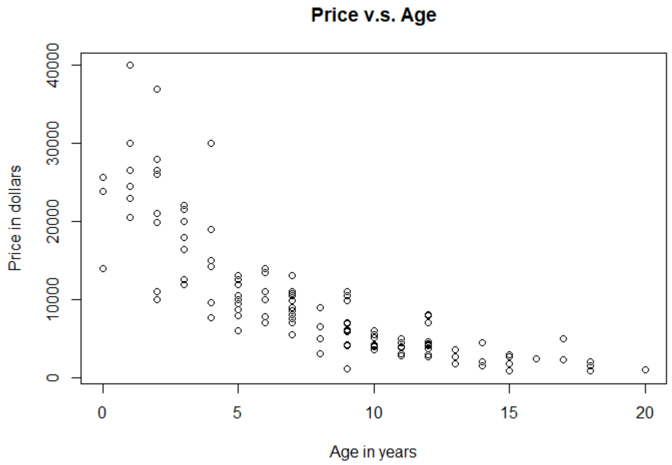
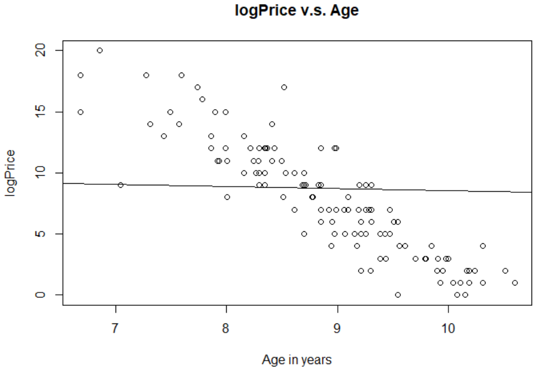
CAR PRICES
Xinyi (Anny) Cui

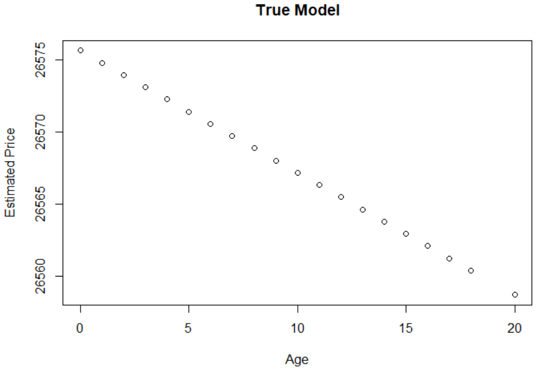
a. After plotting each variable, the graph of price's density plot appears a clear left skewness. Thus, we can do a log transformation on price.



b. When we plot the original data of price v.s. the age (91-year), the graph displays deceasing concave up scatters. This means the price of cars will decrease with the increase of the age, but decrease more and more slowly.

After plotting the log price v.s. the age, we can find a decreasing trendline from the scatter plot. There seems exists a linear relationship between the logPrice and the age, as the log price will decrease with the increase of the age. While, at this moment, the graph has a linear decreasing trend.



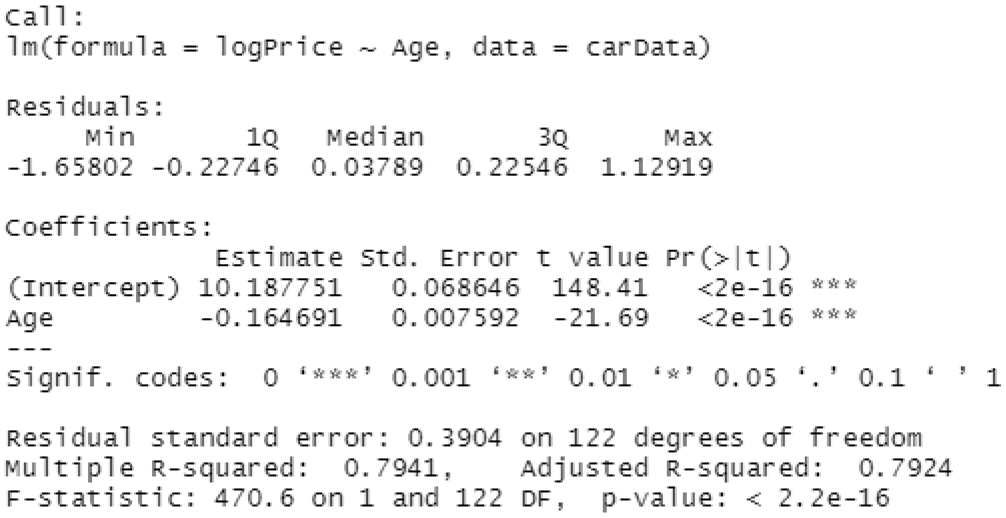


c. The linear model given from our log data is $logPrice = 10.1878 - 0.1647(Age)$. We need to transfer the values of coefficients for the "true" model. This gives a model to estimate the price of cars depends on the age:

$$Price = 26575.66 - 1.179(Age)$$

The number 26575.66 is the constant for intercept, and the number -1.179 means the price of the car will decrease to by 17.9% with one unit of the age increases.

d. As we fit the model by using the log price so we will analyze the summary table with the form of $logPrice = \beta_0 + \beta_1(Age) + \epsilon$. The F-statistic is high which shows a better variability of the model than the constant one. Also, from the t-test, the p-value is $< 2.2e - 16$, and this number is much smaller than 0.05. This means the logPice and Age has a clear linear relationship and there is a trend shown in the graph.



```
Call:
lm(formula = logPrice ~ Age, data = carData)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.65802 -0.22746  0.03789  0.22546  1.12919 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.187751   0.068646  148.41   <2e-16 ***
Age         -0.164691   0.007592  -21.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3904 on 122 degrees of freedom
Multiple R-squared:  0.7941,	Adjusted R-squared:  0.7924 
F-statistic: 470.6 on 1 and 122 DF,  p-value: < 2.2e-16
```

e. The coefficient of determination R^2 is 0.7941. This means 79.41% of the variation in the logPrice can be explained by the data of the age. This is a pretty high percentage and displays a goodness of fit of this model and so that we can assume when we transform the coefficient with the function $exp()$, the “true” model will also fit the given values well in the fitted equation.

f. First, we can do the hypothesis of the t-test for $\beta_1 \neq 0$, which means assuming there exists a real trend for the price responding to the age. From the summary table, we can find that the p-value of the slope (Age) is much less than 0.05. The p-value of slope is $< 2e - 16$, statistically significant, and this proves that there is a real trend between the logPrice and the age.

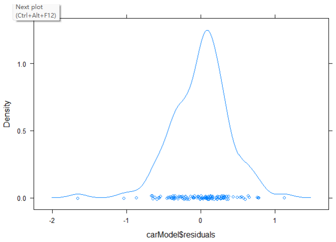
Second, we can do F-test to show the linear model explains more variability than the constant model. The F-statistic compares how much variation is captured by the model and how much is left to the residuals. The value of F-statistic is 470.6 on 1 and 122 DF, and comparing this to an F-distribution with 1 numerator and 122 denominator degrees of freedom, we find an almost 0 p-value. Thus, we can conclude that the logPrice and the age have some relationship and this model based on the age is effective for predicting the price of cars if we convert the log values back.

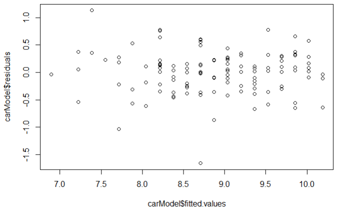
The last test for the hypothesis is checking the correlation coefficient, the R^2 value. We can also do a measure of the strength of the linear relationship by comparing the R value to approach this method. The value of R^2 is 0.7941, which means 79.41% of the variation in the logPrice can be explained by the values of age. Then, we can use the $cor()$ function to find the value of R:

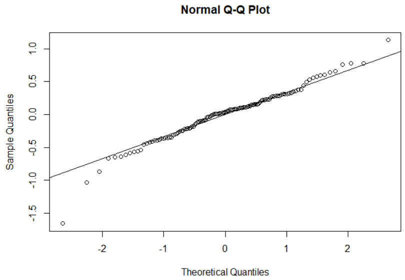
$$\begin{aligned} cor(carData\$logPrice, carData\$Age) &= -0.8911344 \\ cor(carData\$Price, carData\$Age) &= -0.8017428 \\ cor(\text{``}true\text{''}\, Model, carData\$Age) &= -1 \end{aligned}$$

Those negative R values approaching to -1 shows a clear negative strong relationship between the price and the age.

We can also check the linearity of the fitted model by plotting:







The density plot of the residuals show a well normal distribution, the graph of the fitted values v.s. residuals displays an almost zero mean, and the normal quantile graph shows a clear increasing trendline. The linearity of the log model is proved, and then we can assume a good linearity of the model for the one we convert the log values back to the general prices. This is because both response and explanatory variables and even the coefficient of the intercept are done with the $exp()$ function in R to transform back to the general values, and the median number will not change too much after the transformation.

g. From the logPrice fitted model, we can see the confidence interval is $[-0.1797199, -0.1496619]$ for the slope. Then the 95% confidence interval for the "true" model is $[-1.196882, -1.161441]$. The fitted estimation of the slope is -1.179029.

h. From the previous two problems, we can see the linearity of our model which provides a strong negative relationship between the price as a response variable and the age as an explanatory variable after doing the transformation for the linear model. This implies that the price of a Mazda selling in Australia in 1991 will decrease if its age is longer.

i. The 95% confidence interval for the selling price of a 1985 Mazda is $[9184.43, 10656.73]$ and the fitted value for the price is 9893.23 dollars.

j. The 95% prediction interval for the selling price of a 1985 Mazda is $[4551.881, 21502.31]$ and the fitted value for the price is also 9893.23 dollars.