**Problem introduction**

A company offer a maintenance service where for a yearly membership fee; members receive service on call. For each service call, the company incur significant cost. For budgeting purpose therefore, the company would like to predict “Number of Service Calls” each month. The company thinks that number of calls (tot\_calls) each month depends on following variables (mostly related to weather): month, tot\_mem, avg\_temp, deg\_ht, pr\_rain, pr\_snow, age. Use stepwise regression in R to build the “best” Linear Regression model or take an appropriate action to modify the model for prediction.

**Original model**

**Step 1 Data**

Use the pair() function to check whether there are relation between variables (See Addendum)

**Step 2 Estimation**

**Based on the AIC and Stepwise Regression, we get the model:**

**The mainly influenced factors:**

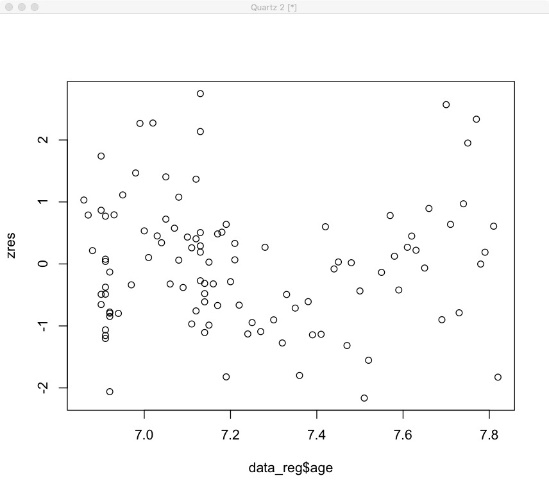
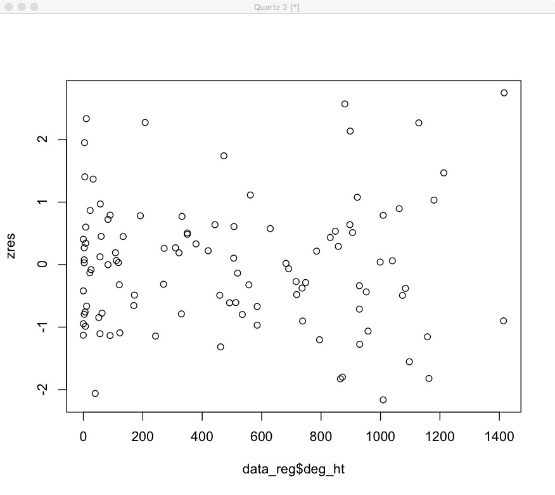
tot\_mem = Total number of members with service contract  
month = Month of the year  
avg\_temp = Average temperature during the month  
deg\_heat = heat index  
pr-rain = total rain during the month  
pr\_snow = total amount of snow during the month  
age = Average age of the equipment registered under the contract

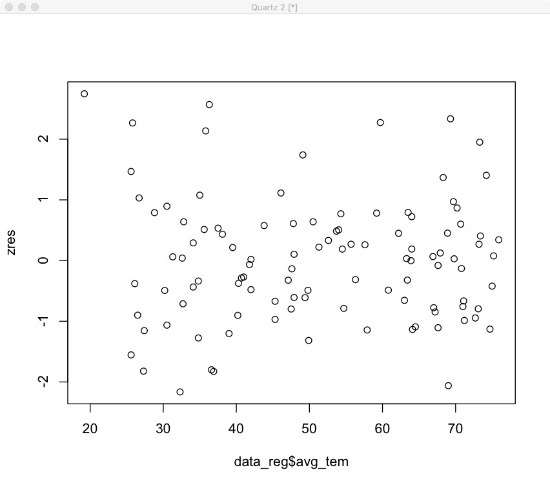
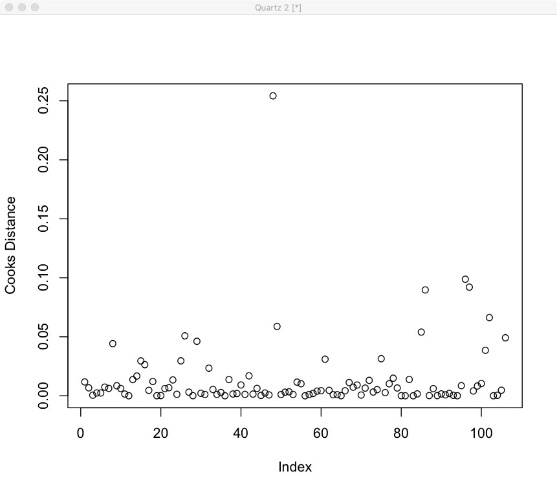
**Step 3 Assessment**

The residual standard error of this model is 1590 so the margin of error for a 95% confidence level is about 2\*1590 = 3180. The adjusted R-squared value is 0.8845, so this model can explain about 88.5% of the variance of service calls.

**Step 4 Testing**

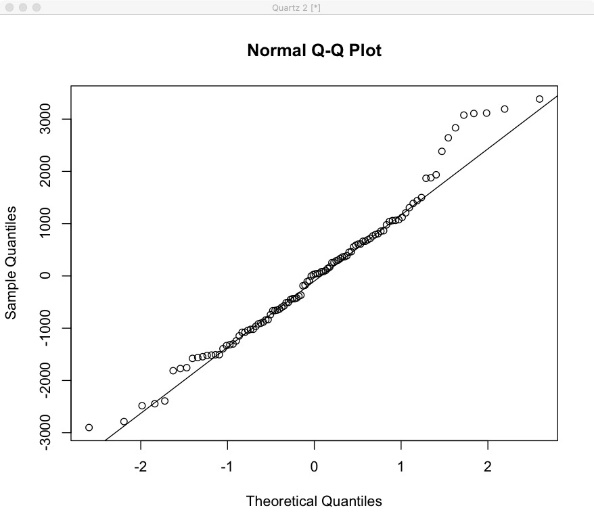
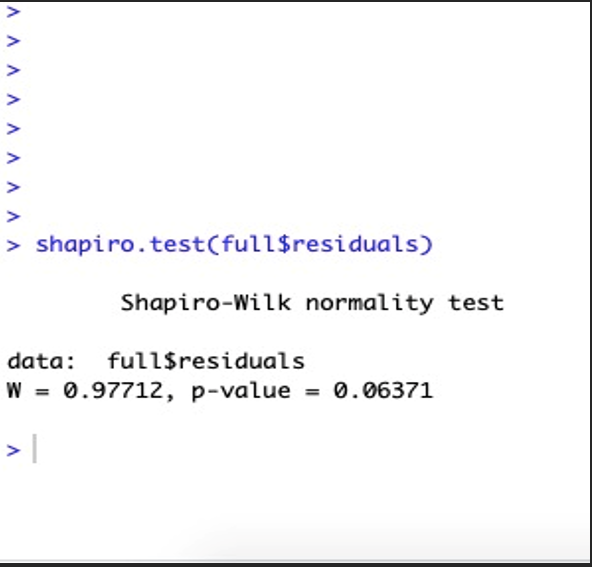
***#Testing Linearity***

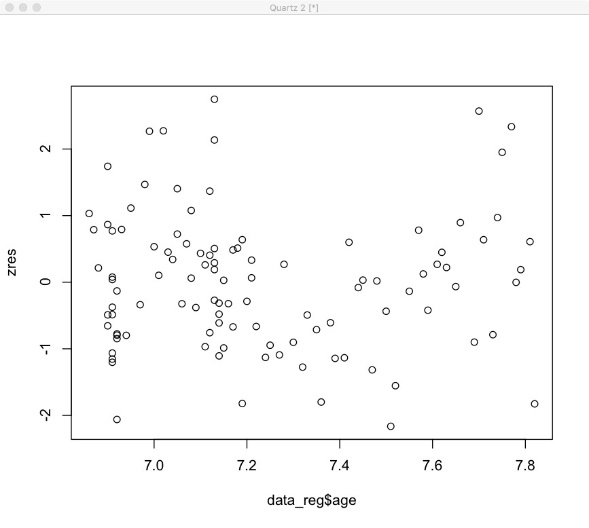
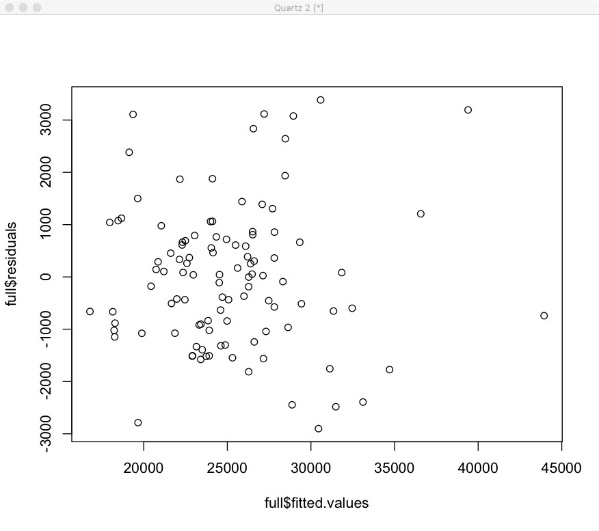
It is hard to tell that the linear assumption is valid, because of some patterns related to x and the last image is outlier image.

***#Test for Normality***

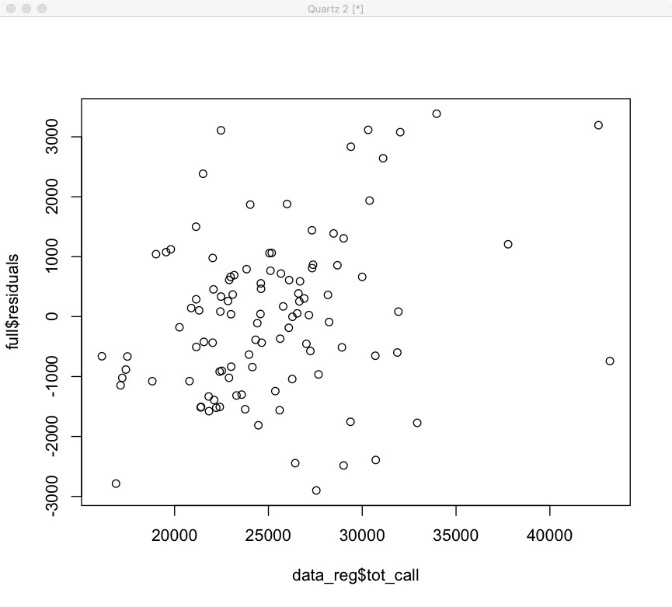
The W-value is 0.97712, which is larger enough and Normality test passed, and P-value is close to Zero, so we need to continue to revise the model.

***#Test for Homoscedasticity***

The homoscedasticity test doesn’t pass, because the first image shows one kind of “U shape” pattern.

***#Test of Independence***



The independence test passes, because the W value is 1.4505, according to the Durbin-Watson Statistics.

**Improve original model**

**The final model:**

We use stepwise regression model to find the optimal model with lowest AIC value. Firstly, fix the factor in the null and full model and then add the new factors into the null model and calculate the AIC value. If AIC value is greater than before, the factor added are less significant to the total calls and if the factor added are more significant to the total calls when AIC value is lower.

**Step 1 Data**

Use the pair() function to check the whether there are relation between variables (See Addendum).

**Step 2 Estimation**

**The mainly influenced factors:**

tot\_mem(Total number of members with service contract)

deg\_ht(Heat index)

avg\_tem(Average temperature during the month)

age(Average age of the equipment registered under the contract)

pr\_snow(Total amount of snow during the month)

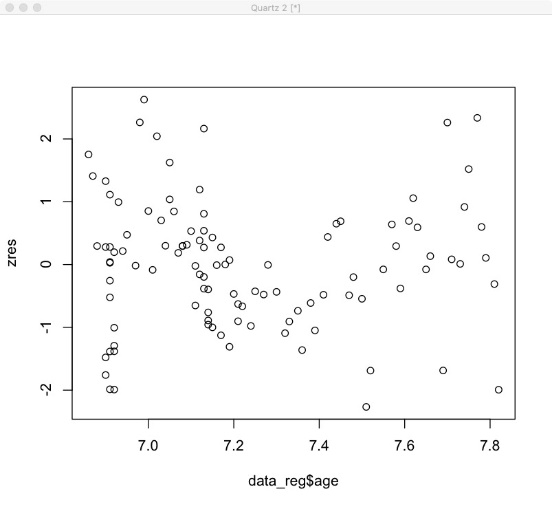
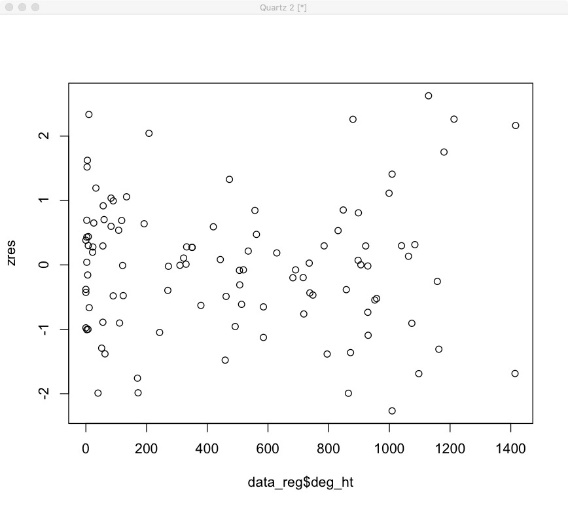
Month(MAR DEC FEB SEP)

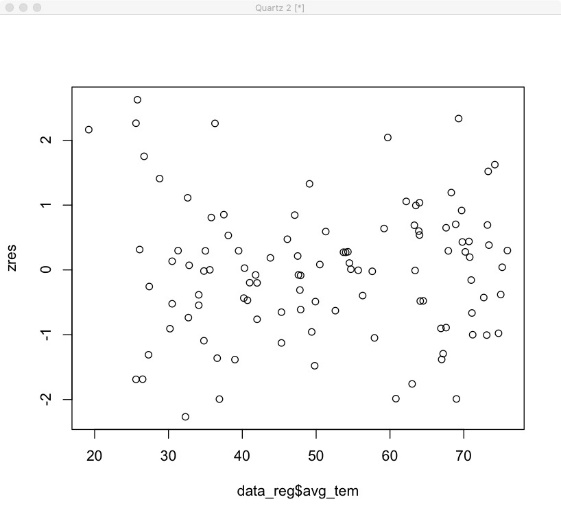
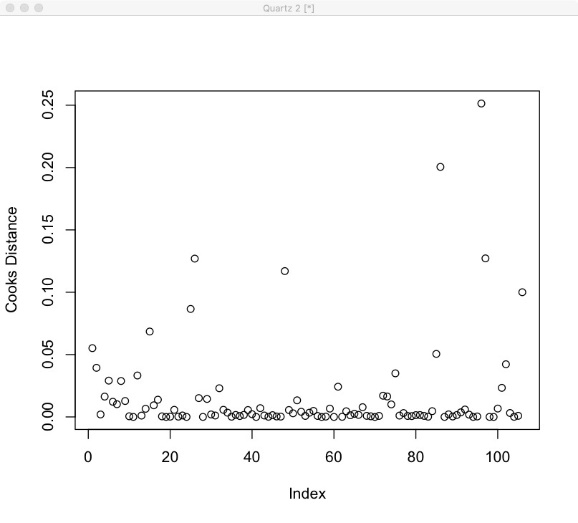
**Step 3 Assessment**

The residual standard error of this model is 1579 so the margin of error for a 95% confidence level is about 2\*1579 = 3158. The adjusted R-squared value is 0.8959, so this model can explain about 89.6% of the variance of service calls.

**Step 4 Testing**

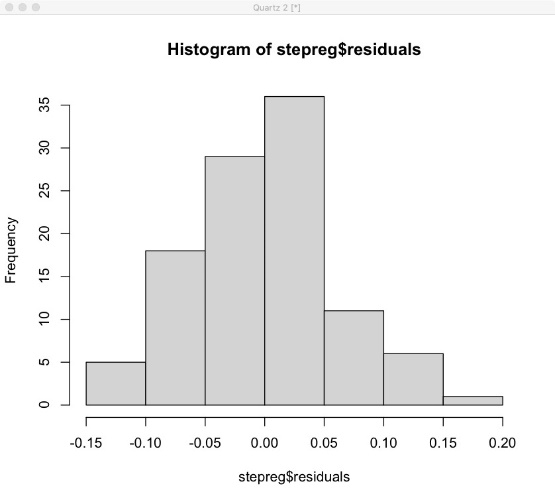
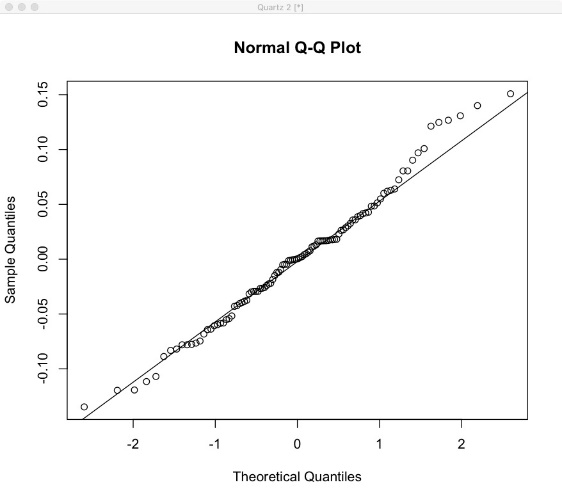
***#Testing Linearity***

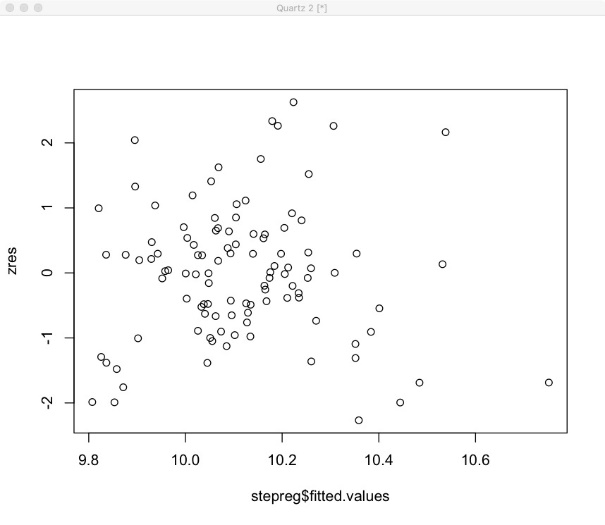
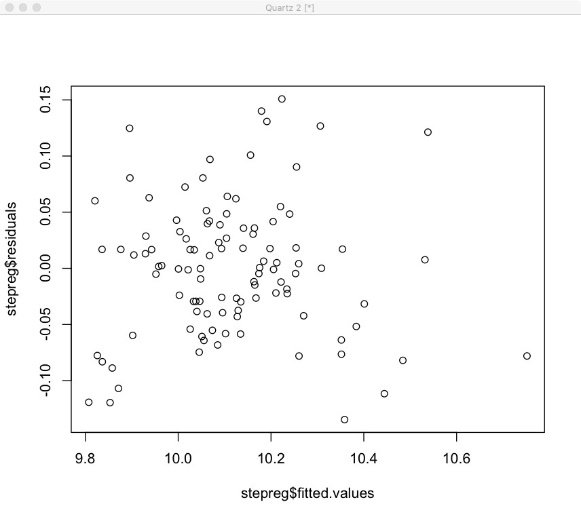
The linearity test passes, according to the plots. In addition, the last image is the outlier image.

***#Test for Normality***

** 

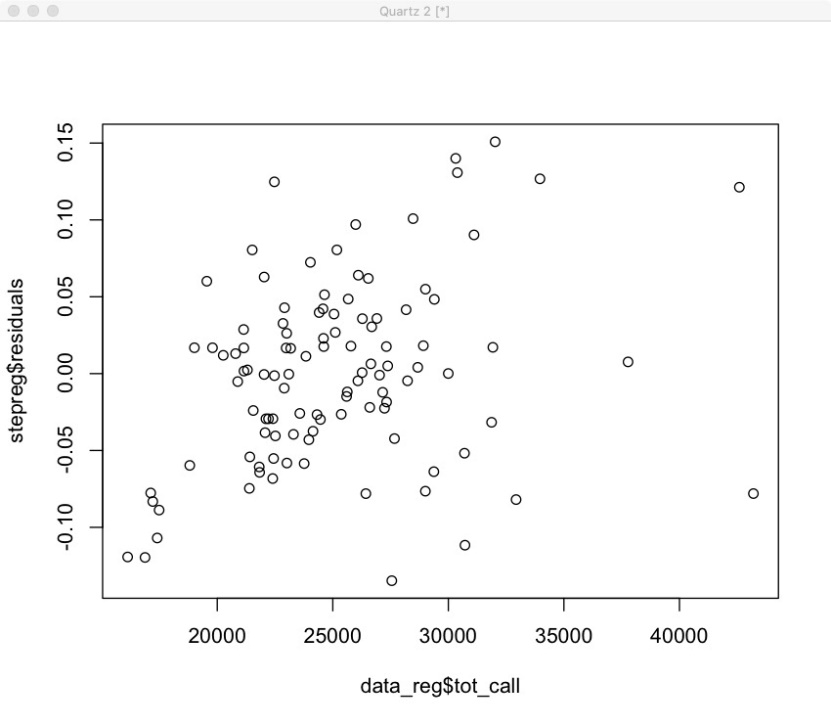
The W-value is 0.98746, which is larger enough and Normality test passed, and P-value is larger than before.

***#Test for Homoscedasticity***

The homoscedasticity test passes, because there is no any pattern in these images.

***#Test for Independence***



The independence test passes, because the W value is 1.1126, according to the Durbin-Watson Statistics.

**Conclusion**

We add the log to the dependent variable(tot\_call) which make the residuals normalized. And we removed some variables with large p-value and got fewer variables. This could be convenient for the company to collect the data. The AIC value is the lowest with applying this model. The adjusted R-squared is larger than the old model and the new model can explain 87.94% of the variance.

**Addendum**

