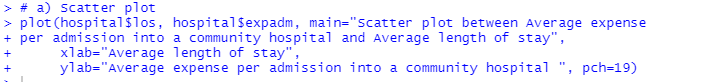
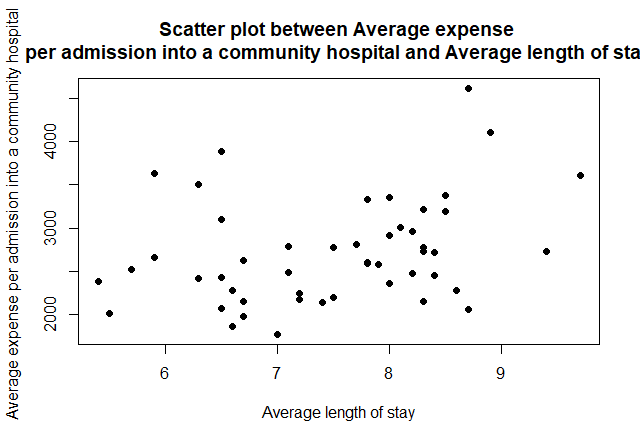
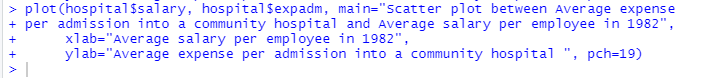
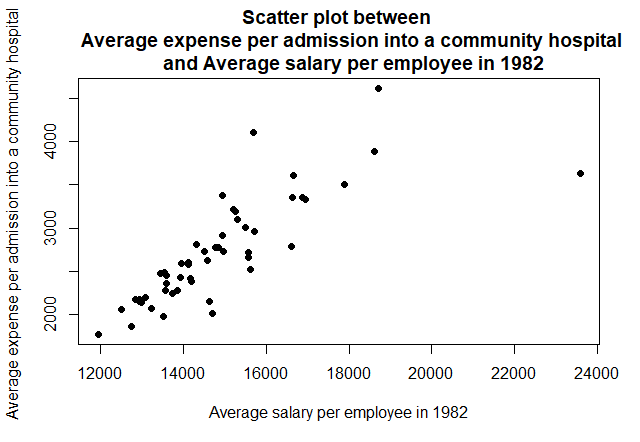
HUYNH NGUYEN

1)

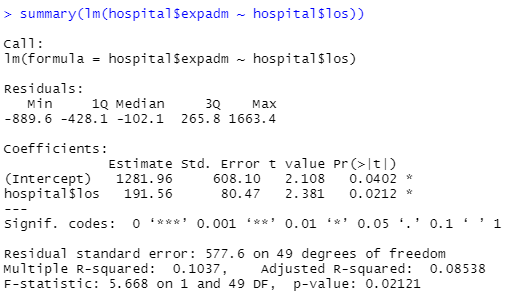
a) 







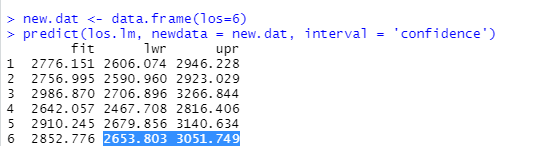
b) Y = 1281.96 + 191.56\*X



c) Suppose X = 6

-> Y1 = 1281.96 +191.56\*(6) = 2431.32

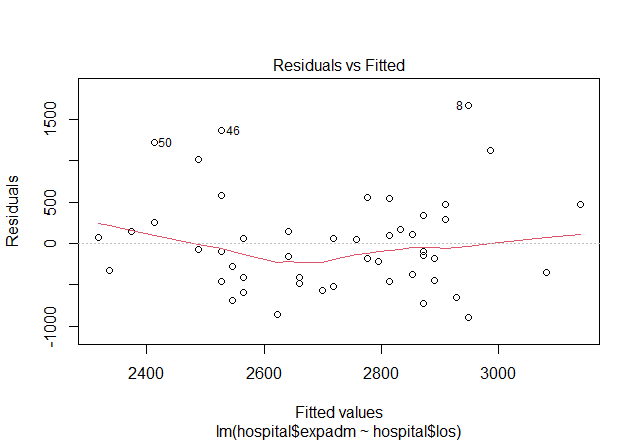




The CI for the average expense per admission at this hospital when average length of stay per admission at Fairyland Community Hospital is 6 days is **(2653.803 , 3051.749)**

d). Since the p-value = **0.0212 <0.05**. Hence, there is sufficient evidence to conclude that there is a significant linear relationship between expense per admission and the length of stay because the correlation coefficient is significantly different from zero.

e). 

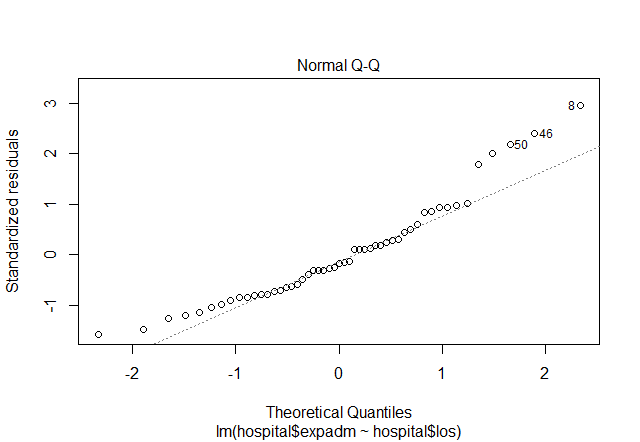


The regression model does not provide a great fit to the observed data. Besides the data scattering around, there are also some outliers which are 8th, 46th and 50th.

The coefficient of determination is 0.1037, so only 10.37% of the variability in average expense per admission into a community hospital is explained by the linear relationship.

The scatterplot shows that there is no evidence of that the assumption of homoscedasticity has been violated.

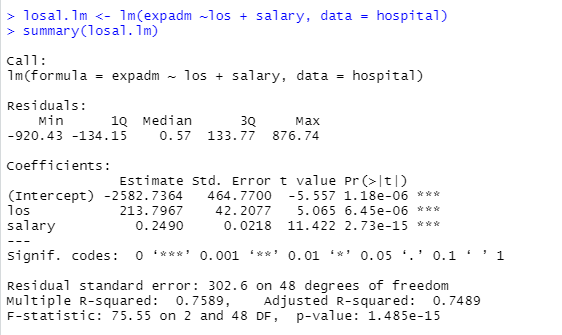




Mostly, the residuals are normally distributed except for some outliers (50th, 46th, 8th )

f). Linear regression model:

Y = -2582.7364 + 213.7967\*(X1) + 0.2490\*(X2)

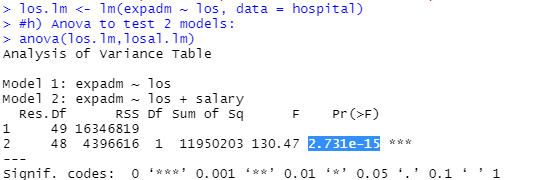


The estimated coefficient of Average length of Stay (los) is 213.7967. This implies that, given the Average Salary (salary) is constant, each day increase in average length of stay causes average expense per admission into a community hospital to increase by 213.7967 on average.

The estimated coefficient of Average Salary (salary) is 0.2490. This implies that, given the length of stay is constant, each one-unit increase in salary average expense per admission into a community hospital to increase by 0.2490 on average.

g) When average salary is added to the model, the estimated coefficient of length of stay increases from 191.56 to 213.7967

h) After Salary is added to the model the R-squared increases from 0.1037 to 0.7589 which means the variability in average expense per admission into a community hospital is explained by the linear relationship has been increased from 10.37% to 75.89%. The coefficient of Salary is statistically significant (p-value = 2.73e-15). Both indicate that salary should be used. An F-test is conducted to confirm this.



Since the p-value is 2.731e-15< 0.05. We conclude that the inclusion of salary in addition to average length of stay significantly improve the ability to predict mean expense per admission.

Problem 2:

Objective: To define whether there is statistically significant difference in the item prices between OldAccount and NewAccount of a student.

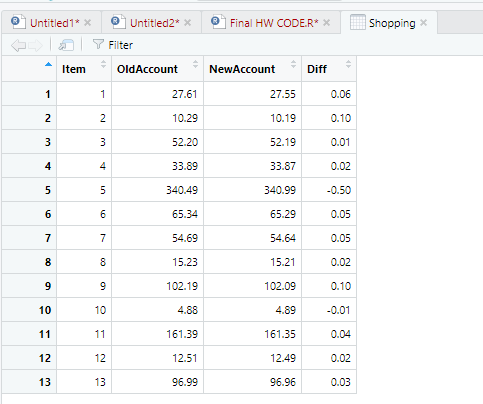
Summary of data:

There are 13 pairs of item prices. The type of prices is set as continuous.

Besides, the 2 samples (OldAccount and NewAccount) are from the populations which are not independent (Same items but are viewed from different account)

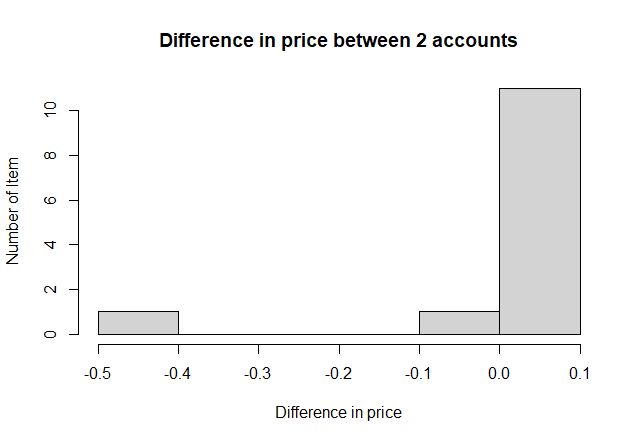
Assumptions:

The 2 populations are not independent because the prices of items are collected from the same e-commerce website with different accounts.



I create a column Diff which equals the difference in prices between the Old and New account.



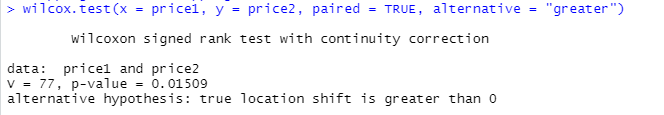


The differences in Table Shopping are displayed in Figure above in the form of a histogram. The graph confirms that differences in reduction in FVC may not be normally distributed (although we cannot say for sure, since the sample size is small); therefore, the paired t-test would not be appropriate. Instead, Wilcoxon Signed-Rank Test will be conducted to test the null hypothesis:

H0: that the median of the underlying population of differences ≤0

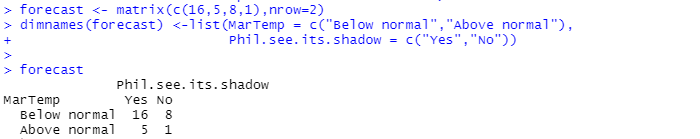
Against

H1: that the median of the underlying population of differences > 0



The p-value is 0.01509 is < 0.05. Thus, I reject the null hypothesis and conclude that the prices from the old account are greater than that of the new account. In other words, there is evidence that the e-commerce company has applied pricing strategy on the items that a student is likely to buy.

Problem 3:



a) To examine the effectiveness Phil forecasting the March Temperature, we wish to know whether there is an association between the below normal March temperature and the times Phil sees its shadow among years which have had below normal March temperature. To determine this, we test the null hypothesis:

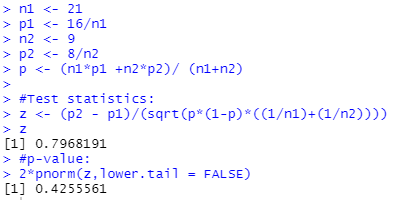
H0: the proportion of years when March temperature is below normal among the population years Phil sees its shadow is identical to the proportion of years when March temperature is below normal among years when Phil doesn’t see its shadow.

Against the alternative hypothesis:

H1: The proportions of years when March temperature are below normal are not identical in the two populations.

b). Although we are considering only one tail of the chi-square distribution, the test is two-sided; large outcomes of can result when the observed value is greater than the expected value and also when it is smaller. However, as a rule of thumb, the Chi square test is an approximate test, we may use it whenever each cell has ≥ 5 observations. Since there is 1 cell has just 1 observation. I’d like to use the z-test instead.

c).



The p-value is 0.4255561 which is > 0.05. Hence, we fail to reject the null hypothesis and conclude that the proportion of years when March temperature is below normal among the population years Phil sees its shadow is identical to the proportion of years when March temperature is below normal among years when Phil doesn’t see its shadow. In other words, there is no significant relationship between Phil sees its shadow and March temperature.

d). The p-value is more than 0.05 that means accepting the null hypothesis. Meaning that there is no significant relationship between Phils sees shadow and temperature forecast. That means there is no need to depend on Groundhog Day's information to predict the March temperature.

Problem 4:

