Homework 8

Kevin Gardner

Due 11/02/2016

Use the following query to acquire the data. Export the 74,596 records to JMP.

SELECT \* from carbo\_transactions WHERE upc in (3000005970, 3000005300,  
9999967727) and geography = 1 ORDER BY upc;

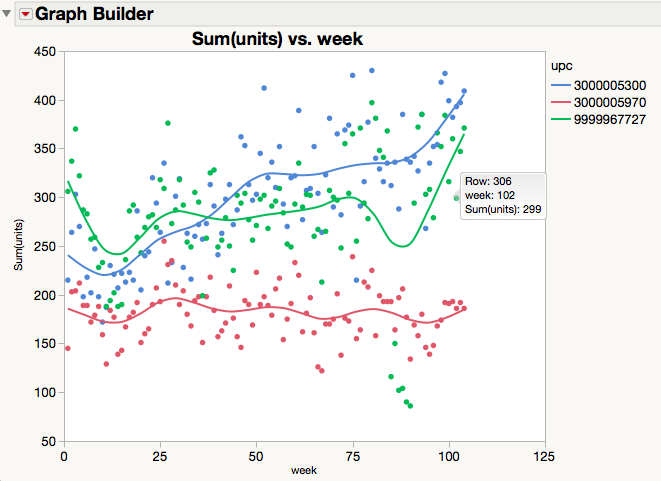
## Question 1

Identify what products these three upc denote. Using Graph Builder, plot the weekly sales volume (number of units) for each of these three products over time. Fit a spline to each. Describe what you see.

Querying the carbo\_product\_lookup for the three upc returns the following result. From this we see that 3000005300, 3000005970 and 9999967727 are Aunt Jemima pancake mix, Aunt Jemima syrup and Private Label syrup, respectively.



Below is a plot of weekly sales volume for each of these three products over time. From this we see that Aunt Jemima pancake mix (3000005300) tends to be the highest selling of the three products and is closely followed by Aunt Jemima syrup (3000005970). Weekly units sold of 3000005970 and 3000005300 are trending upwards over time, whereas the trend is slightly negative for 9999967727.



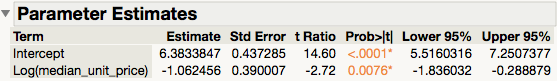
## Question 2

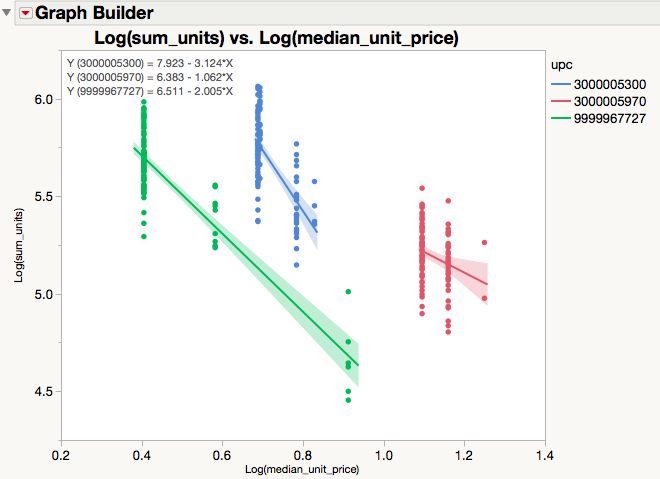
Using median unit price each week has advantages over using mean unit price per week. Answer the following questions using simple regression models with X = ln(weekly median unit price).

1. Does Aunt Jemima syrup weekly unit sales show price sensitivity? Give a 95% confidence interval for price elasticity of demand.

Aunt Jemima syrup weekly unit sales does shows some price sensitivity in that its demand curve has a negative slope. From the graph this is indicated by the slope of the red line, which is less negative slope than the other two slopes.

Interpreting the model for 3000005970 we have that a 1% increase in log median unit price corresponds with about a 1.062% decreases in log sum units sold. We are 95% confident that the price elasticity of demand is between -1.836032 and -0.288879.

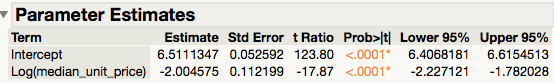




1. Estimate the price sensitivity for the Kroger store brand (Private Label) syrup and compare this result with that for Aunt Jemima syrup. Does the difference make sense?

Interpreting the model for 9999967727 we have that a 1% increase in log median unit price corresponds with about a 2.005% decrease in log sum units sold. We are 95% confident that the price elasticity of demand is between -2.227121 and -1.782028.

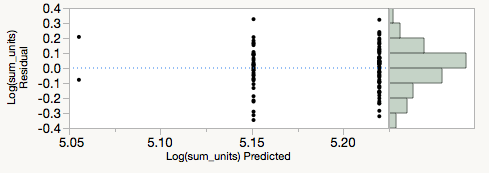
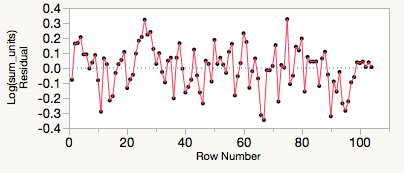
Demand for private label syrup is more sensitive to price than Aunt Jemima syrup and this does not surprise us.

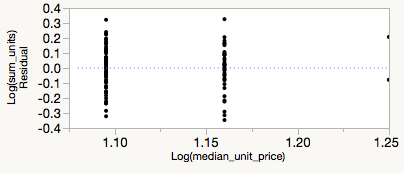
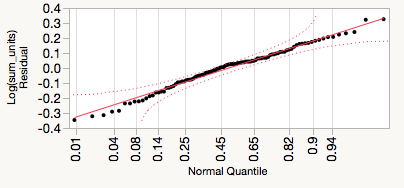


## Question 3

Show residual plots for the regression models fit for 2a and 2b and try to validate whether these models do or do not satisfy the assumptions of constant price elasticity, constant variance, and independently distributed normal errors (residuals).

**Residual plots for 2a - Aunt Jemima syrup (3000005970)**

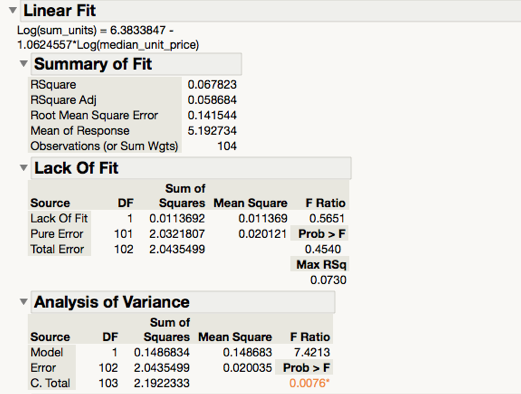




**Constant price elasticity**

From the Residual by Predicted plot we do not see any strong patterns that suggest nonlinearity. The residuals are somewhat smaller when predicted log sum units is 5.05, but this could be due to the fact that there are only two weeks with this level of predicted log sum units.

The linear model assumes constant price elasticity, so to determine if constant price elasticity is satisfied we can use the lack of fit test. The linear model only explains about 7% of the variation in log sum units, but the most complex model does not provide significant improvement. The p-value for lack of fit is 0.4540, so we fail to reject the null hypothesis and conclude the linear model is adequate.



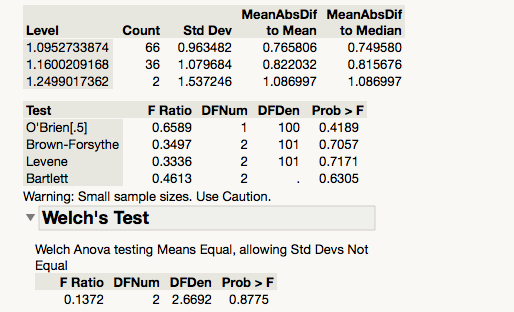
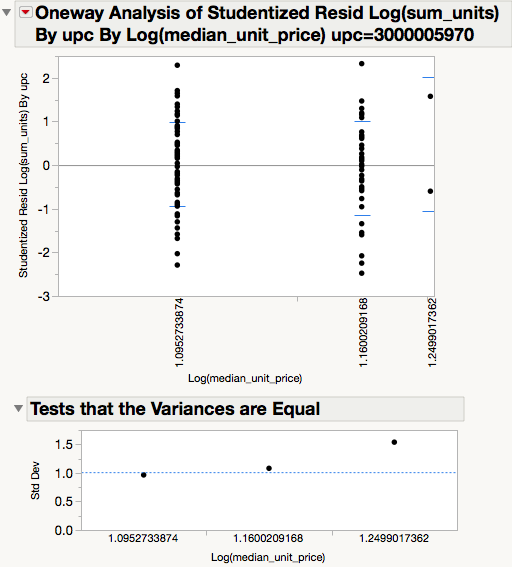
**Constant variance**

From the Residual by log median unit price plot (bottom right) it looks like the residuals have about the same variance for different values of log median unit price.

The residuals are smaller when log median unit price is $1.25, but this could be due to only having two weeks with this level of log median unit price.

In order to test constant variance of the residuals we perform a one-way analysis of Studentized Residual by log median unit price as a categorical variable. The Studentized residuals change slightly with log median unit price and in the second plot we see the standard deviation of the Studentized residual may be increasing with log median unit price.

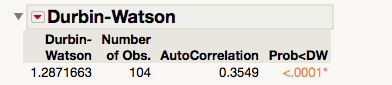
However, all four test tests fail to reject the null hypothesis of constant variance. In particular O’Brien test, which is an ANOVA on the group sample variances, has a p-value of 0.4189. This is not significant so we fail to reject the null hypothesis and conclude that the variances are not unequal.



**Independently distributed normal errors (residuals)**

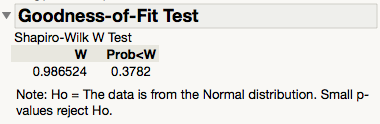
From the Residual by Row Number plot (top left) it looks like positive residuals are more likely to be followed by positive residuals and negative residuals are more likely to be followed by negative residuals, but this is difficult to determine visually.

From the Durbin-Watson test we find the autocorrelation coefficient is 0.3549. The p-value is statistically significant so we reject the null hypothesis and conclude there is positive autocorrelation. This suggests that the errors are not independent.



From the Residual by Normal Quantile plot we do not observe any severe departures from normality. The residual quantiles bend away slightly from the normal quantiles near the tails of the distribution but this does not appear to be systematic.

From the Goodness-of-Fit Test we fail to reject the null hypothesis so can conclude that the errors are approximately normally distributed.



**Residual plots for 2b – Private label syrup (9999967727)**

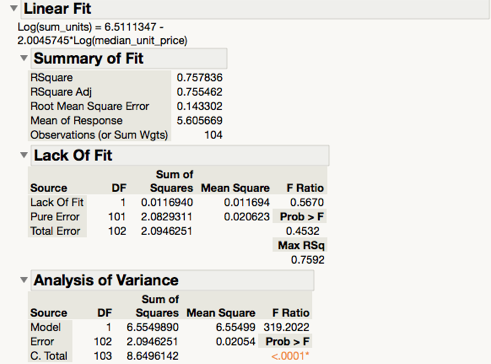




**Constant price elasticity**

From the Residual by Predicted plot we do not see any strong patterns that suggest nonlinearity. The residuals are somewhat smaller when predicted log sum units is 5.3, but this could be due to the fact that there are only two weeks with this level of predicted log sum units.

The linear model assumes constant price elasticity, so to determine if constant price elasticity is satisfied we can use the lack of fit test. The linear model explains about 76% of the variation in log sum units, and the most complex model does not provide significant improvement. The p-value for lack of fit is 0.4532, so we fail to reject the null hypothesis and conclude the linear model is adequate.

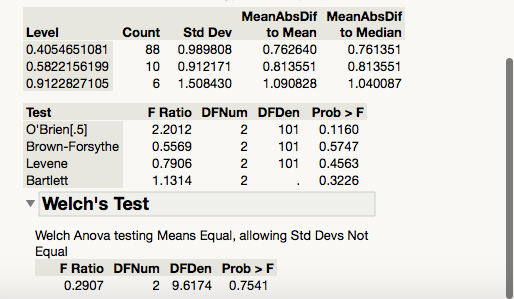
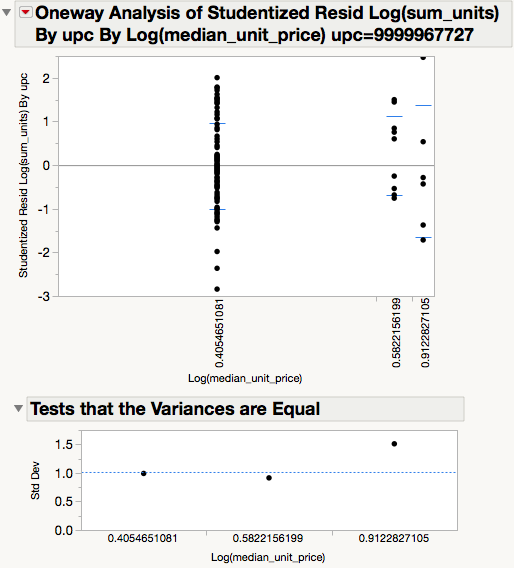


**Constant variance**

From the Residual by log median unit price plot (bottom right) it looks like the residuals have about the same variance for different values of log median unit price.

As in part a, we test constant variance of the residuals by performing a one-way analysis of Studentized Residual by log median unit price as a categorical variable.

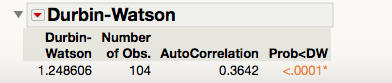
Again, all four test tests fail to reject the null hypothesis of constant variance. The p-value for the O’Brien test (0.1160) is not significant, so we fail to reject the null hypothesis and conclude that sample variances are approximately equal for different levels of log median unit price.



**Independently distributed normal errors (residuals)**

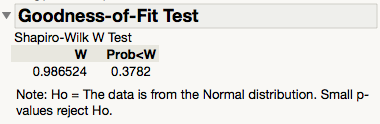
From the Residual by Row Number plot (top left) it looks like positive residuals are more likely to be followed by positive residuals and negative residuals are more likely to be followed by negative residuals, but this is difficult to determine visually.

From the Durbin-Watson test we find the autocorrelation coefficient is 0.3642. The p-value is of <.0001 is significant, so we reject the null hypothesis and conclude that positive autocorrelation is present. This suggests that the residuals are not independent.



From the Residual by Normal Quantile plot we observe some slight departures from normality. The residual quantiles bend away from the normal quantiles near the tails of the distribution, but our sample may be large enough to accommodate these apparent departures.

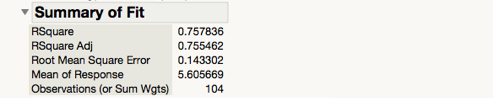
From the Goodness-of-Fit Test we fail to reject the null hypothesis, so we can conclude that the errors are approximately normally distributed.



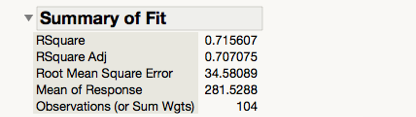
## Question 4

Fit a multiple regression model using Fit Model for Y = ln(weekly Private Label syrup units sold) with X1 = ln(weekly median unit price) and include Week and Week^2 in the models as well. Interpret the following output, explaining how the addition of the terms involving Week.

Including



Not including



1. Changes R2.

Including week and week^2 increased the Rsq from about 0.715607 to 0.757836. So the model explains 4% more of variation in log sum units when we include the terms involving week.

1. Changes the RMSE.

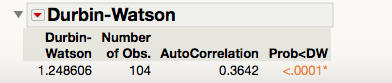
Including week and week^2 decreased the RMSE from 34.58089 to 0.143302. So the absolute fit of the model improves dramatically and (predicted values are closer to observed data points) when we include the terms involving week.

1. Changes autocorrelation in the residuals.

Including

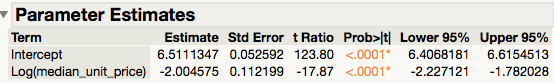


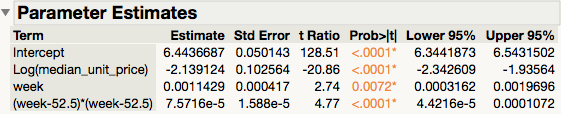
Not including



Including week and week^2 decreased the autocorrelation from 0.3642 to 0.0103. The p-value for the Durbin-Watson test is significant, but it becomes less significant as a result of including the terms involving week.

1. Changes the estimated price elasticity.





Including week and week^2 causes price elasticity increase (become more negative). The coefficient on log median unit price changed from -2.004575 to -2.139124.

## Question 5

The file HW8 sales by household.jmp lists the number of times each household bought these items. Use Analysis – Multivariate Methods – Multivariate to produce a matrix of scatterplots and correlations. Interpret these three coefficients, explaining why they have the sign and magnitude that they do. What would be the effect of adding (0 0 0) for households with no trips on the correlation coefficients?

The pairwise correlation between 3000005970 and 3000005300 is -0.0181. So as a household’s frequency of purchasing Aunt Jamima syrup goes up, its frequency of purchasing Aunt Jamima pancake mix goes down.

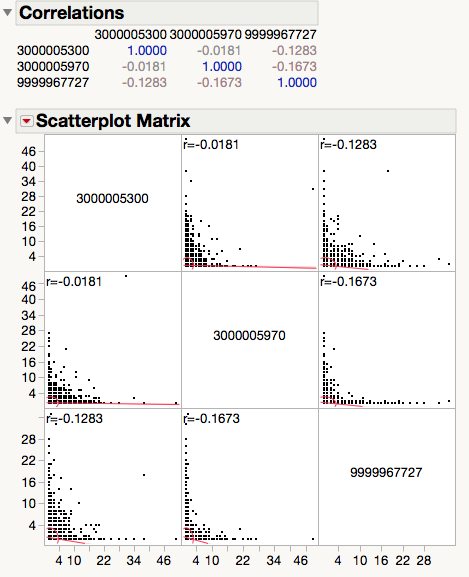
We expected the correlation to be positive, but the unexpected sign is probably due to the sample not being representative. Also, the correlation is weaker and less significant than the other correlations.

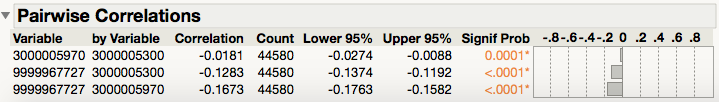
The correlation between 9999967727 and 3000005300 is -0.1283. So as a household’s frequency of purchasing Private label syrup goes up, its frequency of purchasing Aunt Jamima pancake mix goes down.

Similar to the above, we would expect for the correlation between to be positive and the unexpected sign is probably due to the sample not being representative.

The pairwise correlation between 9999967727 and 3000005970 is -0.1673. So as a household’s frequency of purchasing Private label syrup goes up, its frequency of purchasing Aunt Jamima pancake syrup goes down. This correlation is consistent with what we would intuitively expect. This is the strongest correlation in magnitude and significance and this does not surprise us.

Including 0 0 0 for the households who did not purchase either of the three products would provide a more representative sample. For instance, a better sample would be to include all households with purchases in the category of pancake mix and pancake syrup. This would contain observations of 0 0 0 while still allowing us to detect a correlation.





## Extra credit:

Do you see evidence of cross-elasticity for any of these products?

Regressing log sum units 9999967727 on log median unit price 9999967727 and log median unit price 3000005970, we find that cross-elasticity is negative and not significant. The coefficient of -0.149412 suggests that a 1% increase in log median unit price 3000005970 is associated with about a 0.15% decrease in log sum units 9999967727.

Private label syrup and Aunt Jamima syrup are substitutes. So we were expecting the cross-elasticity be positive. Given that it’s negative, we are not surprised that it’s insignificant and the 95% confidence interval contains 0.

