**REGRESSION ANALYSIS HANDOUT**

**B.P.S. Murthi**

The data (soft drinks.xls) provides weekly sales and prices for a large supermarket chain aggregated at each store level, that is, each observation denotes sales of one UPC (universal product code) per week and store. The data cover sales in eight stores and 20 weeks.

The data are somewhat pruned by including sales of only two brands, Brand X and Brand Y, by translating UPC codes into brand names (**BRAND**), and by adding regular versus diet soft drink classifications (**CLASS**). Also included are the following variables:

1. **OBS**: number of each observation
2. **BRAND**: brand name
3. **UPC**: universal product code
4. **STORE**: identification of individual stores in which an observation was recorded
5. **WEEK**: the week in which an observation was recorded

* **OUNCES**: package weight (number of liquid ounces per package)

1. **DEAL**: indicator/dummy of whether a UPC was price-promoted that week (takes on 0 or 1)
2. **FEAT**: indicator of whether UPC was advertised or on in-store display that week (takes on 0 or 1)
3. **HVAL150**: the percentage of real estate property in the so-called store trading area (i.e., the store neighborhood) whose value exceeds $150,000
4. **PRICE**: the price per liquid ounce for a UPC (per week and store)
5. **SALES**: the number of liquid ounces sold for a UPC (per week and store)
6. **NUMCOMP**: the number of competing UPCs
7. **CLASS**: classification as regular versus diet soft drink

Model: MODEL1

Dependent Variable: SALES

Number of Observations Read 1640

Number of Observations Used 1640

Analysis of Variance

Sum of Mean

Source DF Squares Square F Value Pr > F

Model 7 29523558600 4217651229 111.88 <.0001

Error 1632 61522897048 37697854

Corrected Total 1639 91046455648

Root MSE 6139.85778 R-Square 0.3243

Dependent Mean 3286.71220 Adj R-Sq 0.3214

Coeff Var 186.80850

Parameter Estimates

**Parameter Standard**

**Variable Label DF Estimate Error t Value Pr > |t|**

Intercept Intercept 1 6834.17276 2177.93790 3.14 0.0017

PRICE PRICE 1 -170964 80633 -2.12 0.0341

DEAL DEAL 1 11894 5006.64513 2.38 0.0176

FEAT FEAT 1 10416 895.24188 11.64 <.0001

HVAL150 HVAL150 1 1266.09072 1171.46147 1.08 0.2800

NUMCOMP NUMCOMP 1 -37.15207 37.45954 -0.99 0.3214

CLASS1 CLASS1 1 722.50080 312.76359 2.31 0.0210

1. *Comment on the model fit ( R2)?*

R2 =0.32 implies that all the 6 explanatory variables explain 32% of the variance in the dependent variable MOVE (i.e., sales volume). R2 lies in a range between 0 and 1 with values closer to 1.00 indicating a very good model fit. However, in practical applications, since there are many factors that could affect sales, and managers have data on only a few variables, R2 =0.30 and above are considered reasonable model fits.

1. *What is adjusted R2? Why do we need it?*

R2 has the property that as you add more explanatory variables to the model, it increases and could in the limit become equal to 1. For example, in a dataset that has 100 observations, if a researcher uses 99 variables to explain sales, R2 will be equal to 1. So one could get a spuriously high R2, which may be misleading. Adjusted R2 imposes a penalty on any variable added to the model that has a very small explanatory power. Thus, as we add more variables to the model, Adjusted R2 could actually go down and is therefore a more accurate measure of model fit. If the difference between R2 and adjusted R2 is large (i.e., more than 5% points), adjusted R2 is the more accurate indicator of model fit.

1. *What is the hypothesis that t-value is testing? How do you get the t-value?*

It is testing the null hypothesis Ho: β = 0 against the alternate hypothesis Ha: β ≠ 0.

If the t-value is greater than 1.96, then we can reject the null hypothesis with a 95% confidence level. Confidence level indicates that if the test were conducted 100 times, 95% of the time the null hypothesis would be rejected.

First, t-value is obtained by dividing the **coefficient** (or the parameter estimate) by the **standard error**.

1. *What is the p-value?*

‘Pr > |t|’ is also called the **p-value** and is equal to (**1- confidence level**). So if you need a confidence level greater than 95%, p-value should be less than 0.05 (i.e., 1- 0.95).

1. *Which coefficients are significant?*

All coefficients that have a t-value greater than 1.96 (or a p-value less than 0.05) are considered to be statistically significant at the 95% confidence level. That implies that the corresponding variable has a non-zero impact on the dependent variable.

Caution: One need not be very strict about the interpretation of significance. For instance, if the t-value were 1.90, by the above rule we would infer that the variable is not significant. But the reality is that we are 94% confident that there is an effect, but we are not 95% confident. A practical manager would consider 90-95% confidence to be acceptable and so even variables with a p-value =0.10 are considered significant. Whether to use 90% or 95% confidence level, depends on how critical is the decision to be made based upon the analysis.

1. *What’s the effect of a dollar increase in price on MOVE?*

A $1 increase in price would decrease the sales by 170964 units. This is inferred from the negative sign on the coefficient, which suggests that as the price increases, sales would decrease.

1. *What is the effect of a DEAL on MOVE?*

Note that DEAL is a variable that takes the value 1 or 0 depending on whether there is a promotion in the store or not, during that week. Such a variable is called a dummy variable. So the interpretation of the coefficient of a DEAL is that when there is a price promotion, sales would increase by 11894 units.

1. *What is the price elasticity of demand?*

Price elasticity of demand is defined as the percentage change in quantity corresponding to a percentage change in price.

PE = =  = β\*

where Δ indicates the change in either quantity (*q*) or price (*p*).

To compute average estimate of price elasticity, multiply the price coefficient with the average price and divide by the average sales.

1. *Is there a synergy between a price reduction and DEAL? How will you test this?*

To test whether two actions jointly produce a synergistic effect, we include an interaction term into the regression model. For instance, it is common for managers to reduce the price and put up a display which proclaims that the item is on special that week. When these two actions are jointly executed, we should expect the effect of a price decrease and the effect of a DEAL on sales. But usually we get an additional sales boost over and above the sum of the above two effects and this is called interaction between two variables. To test whether there is such an interaction effect, we include a new term (called p\_d here), which is the product of price and DEAL (price\* DEAL) into the regression model, in addition to the original price and DEAL variables already present in the model and run the regression analysis again. We get all the coefficients again, but we are reporting only the coefficient of the interaction term below.

**Parameter Standard**

**Variable Label DF Estimate Error t Value Pr > |t|**

p\_d p\*d 1 -519268 195577 -2.66 0.0080

The coefficient of p\_d is -519268 and is significant. That means that there is an interaction effect or a synergistic effect.

Note the interpretation of the negative coefficient is a sales increase, consistent with the negative coefficient of price (as price is reduced sales would increase).

1. *How will you use the model to predict sales?*

If the values of the explanatory variables are given, one can use the coefficients of the model to compute the predicted sales. So the model can be used to do a what-if analysis for the manager. This facilitates good decision making.

**Assumptions of linear regression model:**

Functional form: Yi = ∂ + βχi+ ei i = 1, …,n

Zero mean of error: E (εi) = 0 for all i

Homoscedasticity: Var (εi) = σ2 for all i

If violated creates a problem of **heteroscedasticity**

Non auto correlation: Cor (εi, εj) = 0 if i ≠ j

Uncorrelated regressor + error: Cor (χi, εj) = 0 for all i ≠ j

If violated creates a problem called **multicollinearity**

Normality: εi  ~ N (0, σ2)

**Multicollinearity**

1. Estimated variances of β become large => low t values

=> do not reject null hypotheses.

1. Signs of coefficients could be incorrect.
2. Results are sensitive to deletion of a single row.

*Sources of multicollinearity*

1. High correlations between independent variables (Xi) are one source of collinearity

2. Absence of high correlation does not imply there is no problem of collinearity. Multicollinearity can also occur if an independent variable can be expressed as the weighted sum of other independent variables.

*Detecting whether there is multicollinearity*

1. Check all pairwise correlations between the independent variables.

2. Look at VIF (variance inflation factor). If VIF is greater than 10 there could be multicollinearity

3. Use COLLIN option in PROC REG. Multicollinearity is present if

1. The condition index is high (more than 100) and
2. Two or more variables have a high proportion of variance

*Solutions*

1. Delete one of the two independent variables
2. Combine the two correlated variables.
3. Tip: Sometimes transformation of one of the variables reduces the correlation
4. Use principal components
5. Use ridge regression

**Heteroscedasticity**

*Detection*

1. Plot standardized residuals against predicted Y. If the variance of the residuals increases or decreases with a corresponding increase in predicted Y then there is heteroscedasticity

*Solutions*

1. Transform the Y variable using a log function or a square root function to mitigate the problem. {That is, use log(Y) or sqrt(Y)}
2. Use weighted least squares (WLS)

Plot standardized residuals against each of the X variables. This tells us two things:

1. Whether the model is non-linear with respect to this variable
2. whether a particular variable is causing the variance of the residuals to increase or decrease. This variable is a good candidate for using in the WLS.