Ran a regression model to explain factors that affect BV003 (Number of active Premium Channels). My dependent variable is BV003.

**Answer: R-Square** = 0.29 implies that all the 13 explanatory variables explain 29% of the variance of the dependent variable BV003(No of Active premium channels). The rest 70% is unexplained variation which might be dependent on other factors/variables which we don’t know about.

**Adjusted R-Square** = Adjusted R-Square in this case is very close to the R-Square that is.2988. Which implies that all the predictors affect the variation in No of Active premium channels. Adjusted R square tells us that though we have 13 independent variables they all might not contribute to the 29% variation in premium channels. There may be only few causing the variation which Adjusted R square considers.

**p-values:** By looking at the p-value of the f test which is <.0001 at 95% confidence level we can say that the model is overall statistically significant as the value is less then .05.

If we look at the individual p values of the independent variables all the p values are less than .05 so we can conclude the all the 13 variables are significantly important to the model and we cannot reject any of them.

**t-values:** The t value tests the null hypothesis of the model. If t statistic is greater than the critical value, then the null hypothesis cannot be rejected.

**Coefficients:** The size of the coefficient of independent variable gives the amount of effect it is having on the dependent variable. The positive sign tells the that the effect is increasing and negative sign indicates effect is decreasing assuming all the other variables as constants.

*Effect of Children on Number of active premium channels*

Children is a dummy variable with a value of 0 or 1. So we can say from the coefficient value that presence of children increases the subscription to premium channels by .08.

*Effect of Income on Number of active premium channels*

Income has negative effect on the subscription and that too very negligible. We can say that the subscriptions do not much depend on the income of a person.

*Effect of Reading and Sports on Number of active premium channels*

By looking at their coefficients we can say that they have negative effect on the subscription of premium channels. If the customer is interested in reading the it decreases the subscriptions by .03 and if the customer is a sports fan, then that’s reducing the subscription by .06.

**f-test:** f-test in regression compares the fits of different linear models. Unlike t-tests that can assess only one regression coefficient at a time, the F-test can assess multiple coefficients. It compares a model with no predictors to the model that you specify. The hypotheses for the F-test of the overall significance are as follows:

* **Null hypothesis**: The fit of the intercept-only model and your model are equal.
* **Alternative hypothesis**: The fit of the intercept-only model is significantly reduced compared to your model.

The p value for this regression model is less than .05 and is significant therefore we can reject the null hypothesis and say that our model with predictors is a better model than the no predictors model.

**Standard Error:** The standard error is an estimate of the *standard deviation* of the coefficient, the amount it varies across cases. They can be used for hypothesis testing and constructing confidence intervals. For example, confidence interval for NFL Programming Subscription is constructed as (0.74330 http://www.jerrydallal.com/lhsp/symbol/pm.gif k 0.02515), where k is the appropriate constant depending on the level of confidence desired. For example, for 95% confidence intervals based on large samples, k would be 1.96.

**DF:** It is the degree of freedom which is equal to 1 for all.

1. Is this a good model?

The r square value which is 29% and is below 30% shows that this **model is not a good model** and needs improvement in terms of choosing the independent variables which can explain more variance for the dependent variable.

1. What is the most important variable affecting sales? How did I know this?

The most important variable affecting sales is BV002 that is Number of Receivers that exist in the customer residence address by the amount of .51. It is highest coefficient among all. As the number of receivers increase in a residence the subscription to premium channels also increase. To find out this the standardized estimate was calculated which is the beta coefficient. The beta coefficients are all measured in standard deviation instead of unit variables so that they can be compared to each other.

1. What other variables do I wish to add to the model? Developed a different regression model with additional variables (the goal is to improve adjusted R-square).

I tried adding few variables to the model to improve the adjusted R-Square to 30% to make it a reasonably fit model.

1. **Travel all:** As per my understanding people of who like travelling tend to watch more number of channels on TV as compared non-travelers because they like to see geographic channels/see various places/travel shows which can increase the subscriptions to travelling channels. After performing the regression analysis this variable did affect the adjusted r square but causing a decrease in no of active premium channels.
2. **Buyer Rating:** The values of this variable are from 1 to 4. According to me frequent buyers respond more to the catalog purchase offers ending up subscribing to more number of channels. By adding this variable, it created a negative impact on the dependent variable of .02.
3. **Mortgage Loan Amount:** People who have a big loan amount to pay will try to save money by subscribing to less TV channels. This affected the adjusted r square a lot by bring it to 30%. It is decreasing the no of premium channels by .03.
4. Correlations between the explanatory variables. Which variables are highly correlated (i.e., correlation greater than 0.60).

After performing the correlation analysis, I found correlation up to 55% but none of the variables are highly correlated that is no correlation coefficient is greater than 60%.