**Linear Regression Examples**

1. It is well known that advertising can increase product sales. A manager would like to determine how advertising effectiveness changes with the level of advertising, i.e., as the total spending on advertising increases. Advertising effectiveness is defined as

Advertising Effectiveness = Sales (in $) / Advertising Spending (in $)

The manager has historic data available on quarterly advertising spending and annual sales. What regression model would you suggest to determine how profit varies with advertising spending? Assume that the time periods are otherwise comparable (ignore seasonality and time-trends)

**Answer: One potential model is to the following:**

**Model 1: Ad Effectiveness = A + B \* Ad Spending + error**

**If B is positive, then Ad Effectiveness is increasing in Ad Spending. If B is negative, then Ad Effectiveness is decreasing in Ad Spending. But there is a problem with Model 1. The dependent variable (Ad Effectiveness) is constructed using the independent variable (Ad Spending). As a result the assumption that the error term is independent of all other variables can be violated. In general, it is not a good idea to construct the dependent variable using one or more of the independent variables.**

**To avoid this problem, we can construct a model that has Sales as the dependent variable.**

**Model 2: Sales = A + B \* Ad Spending + C \* Ad Spending^2 + error**

**Note that the Model 2 considers that Sales is affected by Ad Spending as well as other unobserved factors (the error term) that are independent of advertising, which is a reasonable approximation. Whereas in Model 1 where we use Ad Effectiveness as the dependent variable, we are essentially dividing the effect of the unobserved factors on Sales (the error term in Model 2) also by Ad Spending – as a result the error term in Model 1 includes not only the effect of these unobserved factors on Sales but also the effect of Ad Spending. As a result, the error term in Model 1 is no longer independent of Ad Spending.**

**The manager can use the estimated relation in Model 2 to determine how Ad Effectiveness (Predicted Sales from Model 2 divided by Ad Spending) varies with Ad Spending.**

**Note that if we did not include the quadratic term for Ad Spending in Model 2, then using the predicted sales from the estimated model, we have**

**Ad Effectiveness = (A + B \* Ad Spending) / Ad Spending = A / Ad Spending + B**

**which is always decreasing in Sales (‘A’ captures Sales when Advertising is 0, which would still be positive). So if we had estimated a linear relationship, the model is not flexible enough to allow for the possibility that Ad Effectiveness can increase with Ad Spending.**

**Whereas by including the quadratic term for Ad spending, we have**

**Ad Effectiveness = (A + B \* Ad Spending + C \* Ad Spending^2) / Ad Spending**

**= A / Ad Spending + B + C \* Ad Spending**

**which can be increasing or decreasing in Sales depending on the relative magnitudes of A and C.**

1. A shoes store manager wants to compare the effect of coupons on sales on men and women. He thinks that coupons are more effective in increasing sales for women compared to men. The manager has data on sales with and without coupons for men and for women, How should the manager compare the coupon effectiveness between men and women?

**Answer: In this case the Sales data is from two separate market segments - men and women. Let us first think whether we can do a simple comparison between the segments using a t-test. For example, can we simply compare the sales with coupons between the two segments? This can be done using an independent sample t-test – where the two samples are the sales with coupon for men and for women. However, this does not account for the fact that the sales could have been different between the segments even without coupons. So we cannot fully attribute the difference to the difference in effectiveness of coupons.**

**We can do a simple t-test to determine whether the increase in sales with a coupon for a particular segment (say men) was higher or lower than a particular number. This would also be an independent sample t-test – with the two samples being sales with and without coupons. One might think why not calculate the average increase in sales with coupon for women (by taking the difference in average sales with coupon and without coupon) and then use this to do a hypothesis test whether the increase in sales for men was higher than this number. However, this ignores the fact that our measure of how much coupons increased sales for women is just one estimate coming from one particular sample of women. The actual effect of coupons for women could be a bit higher or lower than what we observed for this sample.**

**So to account for these different problems, we can use the following regression model:**

**Sales = A + B\* Women + C\*Coupons+ D\* Women\*Coupons + error**

**In this model, Sales could be Sales from men or women, with or without coupon – so the sales data from different segments is in one variable. Each observation corresponds to one sale observation from men / women and with / without coupon. Women is an indicator variable to indicate whether the segment (for that particular sales observation) is women or not, and Coupons is an indicator variable to indicate whether a coupon was used or not (for that particular sales observation).**

**In this model, B captures any difference in sales between men and women even without coupons (this is the difference in sales between men and women when the indicator variable Coupons = 0). C captures the difference in sales with and without a coupon for men (this is the difference in sales with and without a coupon when the indicator variable Women = 0). Finally, D captures the difference in effectiveness of coupon between men and women. If D is positive, coupons are more effective in increasing sales for women than men (controlling for all other differences between the segments – which is captured by B).**