IEOR 242 HW1 Solution

Hyungki Im

Sep 2021

Problem 1(15 points)

(a) More complicated model would have lower training RSS. So in this case, cubic regression model has lower training RSS than linear regression model. (Even though the true relationship between X and Y is linear, cubic regression model's training RSS would be lower than the linear regression model's training RSS because of the existence of noise ϵ.)

(b) Since the true relationship between X and Y is linear, we can expect that linear regression model has lower test RSS than cubic regression model's test RSS.

(c) Same answer as (a).

(d) We don't have enough information. The result will largely depends on the test data set.

Grading Rubrics

Each sub question follows this grading rule.

- (-1) Wrong Answer
- (-2) Wrong Justification.

Problem 2(10 points)

$$\hat{y}_i = x_i \times \frac{\sum_{i'=1}^{n} (x_{i'} y_{i'})}{\sum_{j=1}^{n} x_j^2}$$

$$\hat{y}_i = \sum_{i'=1}^{n} \frac{(x_{i'} y_{i'}) \times x_i}{\sum_{j=1}^{n} x_j^2}$$

$$\hat{y}_i = \sum_{i'=1}^{n} \left(\frac{x_i x_{i'}}{\sum_{j=1}^{n} x_j^2} \times y_{i'} \right)$$

$$a_{i'} = \frac{x_i x_{i'}}{\sum_{j=1}^{n} x_j^2}$$

Grading Rubrics

- (-2) Wrong $a_{i'}$.
- (-1) For each calculation mistake.

Problem 3(75 points)

(a): Total 25 points

There is no fixed answer for this question. However, you should have reasonable explanation for your model. In our case, I went in following order

- Drop **Unemployment** as it has high p-value. (model2)
- Drop **CPIAll** as VIF is larger than 5. (model3)
- Drop **CPIEnergy** as p-value is greater than 0.05. (model4)

I refer model1 as a model with all the four independent variable. To my point of view, model2 looks best among these 4 models that I have created. First of all, it has high R^2 value compared to other models and p-value of all the indenpendent variables is stable. Although VIF value of some independent variables is greater than 5, we ignore this information as models with VIF information perform worse than model2. Our linear regression model would be like RogueSales = 1168 + 188.9291*RogueQuries -63.7953*CPIEnergy + 401.8088*CPIAll. Also, coefficient of this linear regression model is reasonable. Model2's R^2 value is 0.795, so we can say that our model performs well with training set observations.

Also, sign of the coefficients make sense. If consumer index of energy price is high, then maintenance cost of vehicle would be high and this might lead to decreased Rogue sales. Also, higher consumer index of all prices(**CPIALL**) might lead to the increase of Rogue sales, as higher **CPIALL** implies economic growth. Finally, higher **RogueQuries** can be considered as higher interest in this product. So positive coefficient seems reasonable.

Grading Rubrics

- (-5) Wrong split of train and test data set.

- (-5) Not explicitly showing final regression model. Representative names for the variables together with coefficients should be presented.(including R^2 value)
- (-5) Considered VIF value while constructing the model.
- (-5) Considered p-value of each variable while constructing the model.
- (-5) Missing interpretation of coefficients. Also, if the interpretation seems unreasonable, deduct 3 points.

(b): Total 15points

The linear regression model would be $RogueSales = -82160 - 976x_1 + 2724x_2 + 3990x_3 + 2682x_5 + 605x_6 + 1713x_7 + 3280x_8 + 689x_9 + 179x_{10} - 255x_{11} + 976x_{12} - 266 * Unemployment + 111 * RogueQueries - 78 * CPIENERGY + 486 * CPIALL$. x_i denotes binary variables for $i-th$ **MonthFactor**. We can see that **MonthFactor** is one of the important variables, as some of the month has high coefficient value. Also **CPIAll** and **RogueQueries** are important for the same reason. We can interpret **MonthFactor** dummy variables as extra sales in a given month compared to baseline month(April in our model) keeping other variables fixed/constant.

Our R^2 is 0.851. So adding the independent variable 'MonthFactor' improves the quality of our model since R^2 value increased from the model of question (a). Answer for *iv*) is in **Grading Rubrics**.

Grading Rubrics

- (-2) Not explicitly showing final regression model.
- (-3) No interpretation of the coefficients of each of the **MonthFactor** dummy variable
- (-3) Missing R^2 of the new model.
- (-2) Missing comparison with previous model.
- (-5) Not presented alternative to model *seasonality*. For example, we can group the **MonthFactor** into **SeasonalFactor** as not all the month are important. One way to implement this is to combine 'Nov',"Dec','Jan' to a seasonal factor 'winter'.

(c): Total 15points

This is an open-ended question. However, you should provide us a reasonable explanation for your chosen model. Here I chosed the model of question (b)

as I considered **Month Factor** is important. Also, I think excluding other continuous variable would not make a significant difference to our model. R^2 value is in our previous answer.

OSR^2 is 0.905 for testing set A, and -0.516 for testing set B (Negative OSR^2 means that the baseline model is performing better than our model.). We can see that our model predicts well until 2019 but predicting worse than the baseline model after 2020. We conclude that some big event(i.e. Corona virus) happened on 2020 and that event has dramatically changed the sales trend of *Rogue*.

Grading Rubrics

- (-3) Not explicitly showing final regression model. If you are using the same model that you have showed us before, then just referring to that model is okay.
- (-3) Missing R^2
- (-3) Missing OSR^2 of testing set A and set B. (1.5 point each)
- (-3) No reasonable explanation for one's model. (Reasonable explanation might be something like: I choose model 1 since it has higher R^2 value compared to model 2)
- (-3) No comparison between OSR^2 value of testing set A and testing set B.

(d): Total 10 points

This is an open ended question. You will be able to get full credits if you get data and build and interpret your model properly.

Grading Rubrics

- (-5) Not using appropriate data. (In most cases, if you present your data this would be fine.) If there is no explanation of the data that has been used, deduct 2 points.
- (-2) Not explicitly showing final regression model.
- (-3) No comparison between your new model and the model on question (c).

(e): Total 10 points

There are two ways to express the loss function. However, they are actually equivalent. Also, you don't have to consider multiple periods in this problem set

since we assumed that the number of units carried over from month to month is always less than or equal to the target inventory levels given by the predictions of your model. However, we will give you full credit's for both cases($T > 1$ and $T = 1$). Let's first see the first solution.

1. When we consider loss function as l_1 = Invertory cost - Revenue.

   Let y_t :=demand of Rogue at time t and $\hat{y}_t$:= expected demand of Rogue at time t. We can express the number of leftovers of Rogue at time t as $max\{0, \hat{y}_t - y_t\}$. The number of Rogue that are sold can be expressed as $min\{y_t, \hat{y}_t\}$. So the profit at time t would be $3000 * min\{y_t, \hat{y}_t\} - 500 * max\{0, \hat{y}_t - y_t\}$. Our loss function can be expressed as

   $$l_1 := -3000 * min\{y_t, \hat{y}_t\} + 500 * max\{0, \hat{y}_t - y_t\}$$

   If we extend our time horizon it would be

   $$\sum_{t=1}^{T} -3000 * min\{y_t, \hat{y}_t\} + 500 * max\{0, \hat{y}_t - y_t\}$$

   . Both answers would given a full credit.

2. When we consider loss function as l_2 = Inventory cost + opportunity cost

   In this case, the expression for invertory cost remains same with previous case. Opportunity cost can be expressed as $max\{0, y_t - \hat{y}_t\}$. So in this case, our loss function would be

   $$l_2 := 3000 * max\{0, y_t - \hat{y}_t\} + 500 * max\{0, \hat{y}_t - y_t\}$$

   If we extend our time horizon it would be

   $$\sum_{t=1}^{T} 3000 * max\{0, y_t - \hat{y}_t\} + 500 * max\{0, \hat{y}_t - y_t\}$$

   Both answers would given a full credit.

Now, let's show that the two formulation are equivalent. If we subtract a constant $3000 * y_t$ on l_2 then we have

$$\begin{aligned} l_2 - 3000 * y_t &= 3000 * max\{0, y_t - \hat{y}_t\} + 500 * max\{0, \hat{y}_t - y_t\} - 3000 * y_t \\ &= -3000 * (min\{0, \hat{y}_t - y_t\} + y_t) + 500 * max\{0, \hat{y}_t - y_t\} \\ &= -3000 * min\{y_t, \hat{y}_t\} + 500 * max\{0, \hat{y}_t - y_t\} \\ &= l_1 \end{aligned}$$

So the two cases are equivalent. Notice that y_t would be a constant in our optimization problem.

Grading Rubrics

- You will get full credits for both $T = 1$ and $T > 1$ cases.
- (-4) -4 points for each missing term in the loss function. (either revenue or opportunity costand,invertory cost)
- (-2) Wrong sign of loss function(didn't change max to min.)