edX (https://www.edx.org)

MITx: 15.071x The Analytics Edge

zhushun0008 (/dashboard) ▼

Courseware (/courses/MITx/15.071x/1T2014/courseware)   Course Info (/courses/MITx/15.071x/1T2014/info)

Discussion (/courses/MITx/15.071x/1T2014/discussion/forum)   Progress (/courses/MITx/15.071x/1T2014/progress)

Syllabus (/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/)

Schedule (/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/)

Help

FORECASTING ELANTRA SALES

An important application of the predictive models studied in this course is to understanding sales. Consider a company that produces and sells a product. In a given period, if the company produces more units than how many consumers will buy, the company will not earn money on the unsold units and will incur additional costs due to having to store those units in inventory before they can be sold. If it produces fewer units than how many consumers will buy, the company will earn less than it potentially could have earned. Being able to predict consumer sales, therefore, is of first order importance to the company.

In this problem, we will try to predict monthly sales of the Hyundai Elantra in the United States. The Hyundai Motor Company is a major automobile manufacturer based in South Korea. The Elantra is a car model that has been produced by Hyundai since 1990 and is sold all over the world, including the United States. We will build a simple model to predict monthly sales using economic indicators of the United States as well as Google search queries.

The file elantra.csv (/c4x/MITx/15.071x/asset/elantra.csv) contains data for the problem. Each observation is a month, from January 2010 to February 2014. For each month, we have the following variables:

- *Month*: the month of the year for the observation (1 = January, 2 = February, 3 = March, ...).
- *Year*: the year of the observation.
- *ElantraSales*: the number of units of the Hyundai Elantra sold in the United States in the given month.
- *Unemployment*: the estimated unemployment percentage in the United States in the given month.
- *Queries*: a (normalized) approximation of the number of Google searches for "hyundai elantra" in the given month.
- *CPI_energy*: the monthly consumer price index (CPI) for energy for the given month.
- *CPI_all*: the consumer price index (CPI) for all products for the given month; this is a measure of the magnitude of the prices paid by consumer households for goods and services (e.g., food, clothing, electricity, etc.).

PROBLEM 1 - LOADING THE DATA (1 point possible)

Load the data set. Split the data set into training and testing sets as follows: place all observations for 2012 and earlier in the training set, and all observations for 2013 and 2014 into the testing set.

How many observations are in the training set?

[   ]   **=**

[   ]

| Check | Save | *You have used 0 of 2 submissions*

PROBLEM 2 - A LINEAR REGRESSION MODEL (1 point possible)

Build a linear regression model to predict monthly Elantra sales using Unemployment, CPI_all, CPI_energy and Queries. Use all of the training set data to do this.

What is the model R-squared? Note: In this problem, we will always be asking for the "Multiple R-Squared" of the model.

[ ]

[ ]

Check   Save   *You have used 0 of 2 submissions*

PROBLEM 3 - SIGNIFICANT VARIABLES

IMPORTANT NOTE: The following problem will not count towards your final grade due to issues we encountered with the auto-grader. You are welcome to still answer it, or to just skip this problem.

Which variables are significant, or have levels that are significant? Use 0.10 as your p-value cutoff. You can select 0-4 of the variables here.

☐ Unemployment
☐ CPI_all
☐ CPI_energy
☐ Queries

Final Check   Save   *You have used 0 of 1 submissions*

PROBLEM 4 - COEFFICIENTS (1 point possible)

What is the coefficient of the Unemployment variable?

[ ]

[ ]

Check   Save   *You have used 0 of 2 submissions*

PROBLEM 5 - INTERPRETING THE COEFFICIENT (1 point possible)

What is the interpretation of this coefficient?

○ For an increase of 1 in predicted Elantra sales, Unemployment decreases by approximately 3000.

○ For an increase of 1 in Unemployment, the prediction of Elantra sales decreases by approximately 3000.

○ If Unemployment increases by 1, then Elantra sales will decrease by approximately 3000; Hyundai should keep unemployment down (by creating jobs in the US or lobbying the US government) if it wishes to increase its sales.

○ For an increase of 1 in Unemployment, then predicted Elantra sales will essentially stay the same, since the coefficient is not statistically significant.

Final Check   Save   *You have used 0 of 1 submissions*

PROBLEM 6 - MODELING SEASONALITY (1 point possible)

Our model R-Squared is relatively low, so we would now like to improve our model. In modeling demand and sales, it is often useful to model seasonality. Seasonality refers to the fact that demand is often cyclical/periodic in time. For example, in countries with different seasons, demand for warm outerwear (like jackets and coats) is higher in fall/autumn and winter (due to the colder weather) than in spring and summer. (In contrast, demand for swimsuits and sunscreen is higher in the summer than in the other seasons.) Another example is the "back to school" period in North America: demand for stationary (pencils, notebooks and so on) in late July and all of August is higher than the rest of the year due to the start of the school year in September.

In our problem, since our data includes the month of the year in which the units were sold, it is feasible for us to incorporate monthly seasonality. From a modeling point of view, it may be reasonable that the month plays an effect in how many Elantra units are sold.

To incorporate the seasonal effect due to the month, build a new linear regression model that predicts monthly Elantra sales using Month as well as Unemployment, CPI_all, CPI_energy and Queries. Do not modify the training and testing data frames before building the model.

What is the model R-Squared?

[   ]

[   ]

[Check] [Save]   *You have used 0 of 2 submissions*

PROBLEM 7 - EFFECT OF ADDING A NEW VARIABLE (1 point possible)

Which of the following best describes the effect of adding Month?

- ○ The model is better because the R-squared has increased.
- ○ The model is not better because the adjusted R-squared has gone down and none of the variables (including the new one) are very significant.
- ○ The model is better because the p-values of the four previous variables have decreased (they have become more significant).
- ○ The model is not better because it has more variables.

[Final Check] [Save]   *You have used 0 of 1 submissions*

PROBLEM 8 - UNDERSTANDING THE MODEL (2 points possible)

Let us try to understand our model.

In the new model, given two monthly periods that are otherwise identical in Unemployment, CPI_all, CPI_energy and Queries, what is the absolute difference in predicted Elantra sales given that one period is in January and one is in March?

[   ]

[   ]

In the new model, given two monthly periods that are otherwise identical in Unemployment, CPI_all, CPI_energy and Queries, what is the absolute difference in predicted Elantra sales given that one period is in January and one is in May?

PROBLEM 9 - NUMERIC VS. FACTORS (1 point possible)

You may be experiencing an uneasy feeling that there is something not quite right in how we have modeled the effect of the calendar month on the monthly sales of Elantras. If so, you are right. In particular, we added Month as a variable, but Month is an ordinary numeric variable. In fact, we must convert Month to a factor variable before adding it to the model.

What is the best explanation for why we must do this?

○ By converting Month to a factor variable, we will effectively increase the number of coefficients we need to estimate, which will boost our model's R-Squared.

○ By modeling Month as a factor variable, the effect of each calendar month is not restricted to be linear in the numerical coding of the month (as demonstrated in Problem 8).

○ Within the data frame, Month is stored in R's Date format, causing errors in how the coefficient is estimated.

Final Check   Save   *You have used 0 of 1 submissions*

PROBLEM 10 - A NEW MODEL (1 point possible)

Re-run the regression with the Month variable modeled as a factor variable. (Create a new variable that models the Month as a factor instead of overwriting the current Month variable. We'll still use the numeric version of Month later in the problem.)

What is the model R-Squared?

[ ]

[ ]

Check   Save   *You have used 0 of 2 submissions*

PROBLEM 11 - SIGNIFICANT VARIABLES (1 point possible)

Which variables are significant, or have levels that are significant? Use 0.10 as your p-value cutoff.

☐ Month (the factor version)
☐ CPI_all
☐ CPI_energy
☐ Unemployment
☐ Queries

Final Check   Save   *You have used 0 of 1 submissions*

PROBLEM 12 - FACTOR VARIABLES IN REGRESSION MODELS (1 point possible)

There are only 11 coefficients provided in the output of the regression model for the Month variable when intuitively there should be 12 (corresponding to the 12 months of the year). In particular, the coefficient for Month = 1 is missing. Why is this?

○ The coefficient for Month = 1 was insignificant and removed.

○ There were no observations in the training set with Month = 1.

○ The effect of Month = 1 is captured in the intercept.

[Final Check] [Save]   *You have used 0 of 1 submissions*

PROBLEM 13 - MULTICOLINEARITY (1 point possible)

Another peculiar observation about the regression is that the sign of the Queries variable has changed. In particular, when we naively modeled Month as a numeric variable, Queries had a positive coefficient. Now, Queries has a negative coefficient. Furthermore, CPI_energy has a positive coefficient -- as the overall price of energy increases, we expect Elantra sales to increase, which seems counter-intuitive (if the price of energy increases, we'd expect consumers to have less funds to purchase automobiles, leading to lower Elantra sales).

As we have seen before, changes in coefficient signs and signs that are counter to our intuition may be due to a multicolinearity problem. To check, compute the correlations of the variables in the training set.

Which of the following variables is CPI_energy highly correlated with? (Include only variables where the absolute value of the correlation exceeds 0.6. For the purpose of this question, treat Month as a numeric variable, not a factor variable.)

☐ Month

☐ Unemployment

☐ Queries

☐ CPI_all

[Final Check] [Save]   *You have used 0 of 1 submissions*

PROBLEM 14 - CORRELATIONS (1 point possible)

Which of the following variables is Queries highly correlated with? Again, compute the correlations on the training set. (Include only variables where the absolute value of the correlation exceeds 0.6. For the purpose of this question, treat Month as a numeric variable, not a factor variable.)

☐ Month

☐ Unemployment

☐ CPI_energy

☐ CPI_all

[Final Check] [Save]   *You have used 0 of 1 submissions*

PROBLEM 15 - A REDUCED MODEL (1 point possible)

Let us now simplify our model (the model using the factor version of the Month variable). We will do this by iteratively removing variables, one at a time. Remove the variable with the highest p-value (i.e., the least statistically significant variable) from the model. Repeat this until there are no variables that are insignificant or variables for which all of the factor levels are insignificant. Use a threshold of 0.10 to determine whether a variable is significant.

Which variables, and in what order, are removed by this process?

○ CPI_energy, then Queries

○ Queries

○ Queries, then CPI_energy

○ Queries, then CPI_energy, then CPI_all

[Final Check] [Save]   *You have used 0 of 1 submissions*

PROBLEM 16 - TEST SET PREDICTIONS (1 point possible)

Using the model from Problem 15, make predictions on the test set. What is the sum of squared errors of the model on the test set?

[   ]

[   ]

[Check] [Save]   *You have used 0 of 2 submissions*

PROBLEM 17 - COMPARING TO A BASELINE (1 point possible)

What would the baseline method predict for all observations in the test set? Remember that the baseline method we use predicts the average outcome of all observations in the training set.

[   ]

[   ]

[Check] [Save]   *You have used 0 of 2 submissions*

PROBLEM 18 - TEST SET R-SQUARED (1 point possible)

What is the test set R-Squared?

[   ]

[   ]

[Check] [Save]   *You have used 0 of 2 submissions*

PROBLEM 19 - ABSOLUTE ERRORS (1 point possible)

What is the largest absolute error that we make in our test set predictions?

[   ]

[   ]

PROBLEM 20 - MONTH OF LARGEST ERROR (1 point possible)

In which period (Month,Year pair) do we make the largest absolute error in our prediction?

- ○ 01/2013
- ○ 02/2013
- ○ 03/2013
- ○ 04/2013
- ○ 05/2013
- ○ 06/2013
- ○ 07/2013
- ○ 08/2013
- ○ 09/2013
- ○ 10/2013
- ○ 11/2013
- ○ 12/2013
- ○ 01/2014
- ○ 02/2014

| Final Check | Save | *You have used 0 of 1 submissions*