Forecasting Elantra Sales (OPTIONAL)

Bookmark this page

IMPORTANT NOTE: This problem is optional, and will not count towards your grade. We have created this problem to give you extra practice with the topics covered in this unit.

Forecasting Elantra Sales (OPTIONAL)

An important application of linear regression is 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 linear regression model to predict monthly sales using economic indicators of the United States as well as Google search queries.

The file 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

0 points possible (ungraded)
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?

Answer: 36

Explanation
You can load the data with the read.csv function:
Elantra = read.csv("elantra.csv")
and then split the data using the subset function:
ElantraTrain = subset(Elantra, Year <= 2012)
ElantraTest = subset(Elantra, Year > 2012)
You can see the number of observations in the training set with the str or nrow function. For the rest of this problem, we will refer to the training set as "ElantraTrain", and the testing set as "ElantraTest".

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

Calculator

Problem 2.1 - A Linear Regression Model

0 points possible (ungraded)
Build a linear regression model to predict monthly Elantra sales using Unemployment, CPI_all, CPI_energy and Queries as the independent variables. 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.

Answer: 0.4282

Explanation
You can build the regression model using the lm function:
ElantraLM = lm(ElantraSales ~ Unemployment + Queries + CPI_energy + CPI_all, data=ElantraTrain)
Then you can find the R-squared value by viewing the model output with the summary function.

Submit You have used 0 of 5 attempts

Answers are displayed within the problem

Problem 2.2 - Significant Variables

0 points possible (ungraded)
How many variables are significant, or have levels that are significant? Use 0.10 as your p-value cutoff.

- 0 ✔
- 1
- 2
- 3
- 4

Explanation
After obtaining the output of the model summary, simply look at the p-values of all of the variables in the output (the right-most column, labeled "Pr(>|t|)"). It turns out that none of them are significant.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

Problem 2.3 - Coefficients

0 points possible (ungraded)
What is the coefficient of the Unemployment variable?

Answer: -3179.90

Explanation
This is the value under the left most column, labeled "Estimate", in the regression output (using the summ
function) for Unemployment.

Calculator

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

Problem 2.4 - Interpreting the Coefficient

0 points possible (ungraded)
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.

Explanation
The second choice is the correct answer; the coefficient is defined as the change in the prediction of the dependent variable (ElantraSales) per unit change in the independent variable in question (Unemployment).
The first choice is therefore not correct; it also does not make intuitive sense since Unemployment is the percentage unemployment rate, which is bounded to be between 0 and 100.
The third choice is not correct because the coefficient indicates how the prediction changes, not how the actual sales change, and this option asserts that actual sales change, i.e., there is a causal effect.
The fourth choice is not correct because the statistical significance indicates how likely it is that, by chance, the true coefficient is not different from zero. However, the estimated coefficient still has a (non-zero) value, and our prediction will change for different values of Unemployment; therefore, the sales prediction cannot stay the same.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

Problem 3.1 - Modeling Seasonality

0 points possible (ungraded)
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 mo

Calculator

the training and testing data frames before building the model.

What is the model R-Squared?

Answer: 0.4344

Explanation
Use the lm function to build the model again, this time with Month included as an independent variable:
ElantraLM = lm(ElantraSales ~ Unemployment + Queries + CPI_energy + CPI_all + Month, data=ElantraTrain)
You can find the R-squared by looking at the summary output.

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

Problem 3.2 - Effect of Adding a New Variable

0 points possible (ungraded)
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.

Explanation
The first option is incorrect because (ordinary) R-Squared always increases (or at least stays the same) when you add new variables. This does not make the model better, and in fact, may hurt the ability of the model to generalize to new, unseen data (overfitting).
The second option is correct: the adjusted R-Squared is the R-Squared but adjusted to take into account the number of variables. If the adjusted R-Squared is lower, then this indicates that our model is not better and in fact may be worse. Furthermore, if none of the variables have become significant, then this also indicates that the model is not better.
The third option is not correct because as stated above, the adjusted R-Squared has become worse. Although the variables have come closer to being significant, this doesn't make it a better model.
The fourth option is not correct. Although it is desirable to have models that are parsimonious (fewer variables), we are ultimately interested in models that have high explanatory power (as measured in training R-Squared) and out of sample predictive power (as measured in testing R-Squared). Adding a key variable may significantly improve the predictive power of the model, and we should thus not dismiss the model simply because it has more variables.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

Problem 3.3 - Understanding the Model

0 points possible (ungraded)
Let us try to understand our model.

Calculator

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?

Answer: 221.38

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?

Answer: 442.76

Explanation
The coefficient for Month is 110.69 (look at the summary output of the model). For the first question, January is coded numerically as 1, while March is coded numerically as 3; the difference in the prediction is therefore 110.69 * (3 - 1) = 110.69 * 2 = 221.38
For the second question, May is numerically coded as 5, while January is 1, so the difference in predicted sales is
110.69 * (5 - 1) = 110.69 * 4 = 442.76

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

Problem 3.4 - Numeric vs. Factors

0 points possible (ungraded)
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. ✔
- Within the data frame, Month is stored in R's Date format, causing errors in how the coefficient is estimated.

Explanation
The second choice is the correct answer. The previous subproblem essentially showed that for every month that we move into the future (e.g, from January to February, from February to March, etc.), our predicted sales go up by 110.69. This isn't right, because the effect of the month should not be affected by the numerical coding, and by modeling Month as a numeric variable, we cannot capture more complex effects. For example, suppose that when the other variables are fixed, an additional 500 units are sold from June to December, relative to the other months. This type of relationship between the boost to the sales and the Month variable would look like a step function at Month = 6, which cannot be modeled as a linear function of Month.
The first choice is not right. As we have discussed before, increasing the number of coefficients will never cause the model's R-Squared to decrease, but if the increase is small, then we have not really improved the predictive power of our model, and converting Month to a factor variable is not justified.
The third choice is also not correct. Month is stored as an ordinary number, so there cannot be any issues due to the Date format.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

Problem 4.1 - A New Model

0 points possible (ungraded)
Re-run the regression with the Month variable modeled as a factor variable. (Create a new variable that models the Month as a factor (using the as.factor function) 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?

Answer: 0.8193

Explanation
To create a new variable that is a factor version of the Month variable, you can use the as.factor function:
ElantraTrain$MonthFactor = as.factor(ElantraTrain$Month)
ElantraTest$MonthFactor = as.factor(ElantraTest$Month)
Then, you want to rebuild the model using the lm function:
ElantraLM = lm(ElantraSales ~ Unemployment + Queries + CPI_energy + CPI_all + MonthFactor, data=ElantraTrain)
You can see the R-squared of the model by looking at the output of the summary function.

Submit You have used 0 of 5 attempts

Answers are displayed within the problem

Problem 4.2 - Significant Variables

0 points possible (ungraded)
Which variables are significant, or have levels that are significant? Use 0.10 as your p-value cutoff. (Select all that apply.)

- [x] Month (the factor version)
- [x] CPI_all
- [x] CPI_energy
- [x] Unemployment
- [ ] Queries

Explanation
Run the summary output of your model and look at the stars/periods on the right.

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Calculator

Problem 5.1 - Multicolinearity

0 points possible (ungraded)

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? Select all that apply. (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 ✔

Explanation
You can use the cor function to compute the correlations:
cor(ElantraTrain[c("Unemployment","Month","Queries","CPI_energy","CPI_all")])
The high correlations between CPI_energy and the other variables are -0.80071881 (Unemployment), 0.8328381 (Queries) and 0.91322591 (CPI_all).

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Problem 5.2 - Correlations

0 points possible (ungraded)

Which of the following variables is Queries highly correlated with? Again, compute the correlations on the training set. Select all that apply. (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 ✔

Calculator

Explanation
You can use the cor function to compute the correlations:
cor(ElantraTrain[c("Unemployment","Month","Queries","CPI_energy","CPI_all")])
Based on these results, we can see that (somewhat surprisingly) there are many variables highly correlated with each other; as a result, the sign change of Queries is likely to be due to multicolinearity.

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Problem 6.1 - A Reduced Model

0 points possible (ungraded)
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

Explanation
The variable with the highest p-value is "Queries". After removing it and looking at the model summary again, we can see that there are no variables that are insignificant, at the 0.10 p-level. Note that Month has a few values that are insignificant, but we don't want to remove it because many values are very significant.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

Problem 6.2 - Test Set Predictions

0 points possible (ungraded)
Using the model from Problem 6.1, make predictions on the test set. What is the sum of squared errors of the model on the test set?

Answer: 190757747

Explanation
First, obtain predictions on the test set by using the predict function:
PredictTest = predict(ElantraLM, newdata=ElantraTest)
Then you can compute the SSE by taking the sum of the squared differences between the ElantraSales variable in the test set and the output of the predictions:
SSE = sum((PredictTest - ElantraTest$ElantraSales)^2)
(Note that for the rest of this problem, we will refer to the test set predictions as "PredictTest".)

Calculator

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

Problem 6.3 - Comparing to a Baseline

0 points possible (ungraded)
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.

Answer: 14462.25

Explanation
The baseline method that is used in the R-Squared calculation (to compute SST, the total sum of squares) simply predicts the mean of ElantraSales in the training set for every observation (i.e., without regard to any of the independent variables).

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

Problem 6.4 - Test Set R-Squared

0 points possible (ungraded)
What is the test set R-Squared?

Answer: 0.7280232

Explanation
You can compute the SST as the sum of the squared differences between ElantraSales in the testing set and the mean of ElantraSales in the training set:
SST = sum((mean(ElantraTrain$ElantraSales) - ElantraTest$ElantraSales)^2)
Then, using the SSE you computed previously, the R-squared is 1 minus the SSE divided by the SST.

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

Problem 6.5 - Absolute Errors

0 points possible (ungraded)
What is the largest absolute error that we make in our test set predictions?

Answer: 7491.488

Explanation
You can get this answer by using the max function and the abs function:
max(abs(PredictTest - ElantraTest$ElantraSales))

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

Calculator

Problem 6.6 - Month of Largest Error

0 points possible (ungraded)

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

Explanation

You can use the which.max and the abs functions to answer this question:

which.max(abs(PredictTest - ElantraTest$ElantraSales))

This returns 5, which is the row number in ElantraTest corresponding to the period for which we make the largest absolute error.

Submit

You have used 0 of 3 attempts

Answers are displayed within the problem

Please remember not to ask for or post complete answers to homework questions in this discussion forum.

edX

Legal

Connect

Calculator