

Homework Assignment #1

August 30, 2018

Problem 1: (10 points)

James Chapter 3, Exercise 4.

Problem 2: (20 points)

Consider the multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon . \quad (1)$$

Let $\lambda_1, \dots, \lambda_p$ be non-zero scalars ($\lambda_j \neq 0$), and consider the *rescaled* features Z_1, \dots, Z_p defined by $Z_j := \lambda_j X_j$. Now consider a different multiple linear regression model based on these rescaled features:

$$Y = \alpha_0 + \alpha_1 Z_1 + \dots + \alpha_p Z_p + \epsilon . \quad (2)$$

Suppose that we have collected data $(x_1, y_1), \dots, (x_n, y_n)$, where each x_i is an observed feature vector of length p and each y_i is an observed scalar response. Similarly as before, define the rescaled observed features by:

$$z_{ij} := \lambda_j x_{ij} \text{ for } i = 1, \dots, n \text{ and } j = 1, \dots, p .$$

Let \mathbf{X} denote the $n \times (p+1)$ matrix whose i^{th} row is $(1, x_{i1}, \dots, x_{ip})$, and likewise let \mathbf{Z} denote the $n \times (p+1)$ matrix whose i^{th} row is $(1, z_{i1}, \dots, z_{ip})$. You may assume that $\text{rank}(\mathbf{X}) = p+1 < n$.

Please answer the following:

- a) (15 points) Let $\hat{\beta} := (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ denote the regression coefficient estimates for model (1), trained on the data $(x_1, y_1), \dots, (x_n, y_n)$. Likewise, let $\hat{\alpha}$ denote the regression coefficient estimates with respect to model (2), trained on the data $(z_1, y_1), \dots, (z_n, y_n)$. Show that $\hat{\alpha}_0 = \hat{\beta}_0$ and

$$\hat{\alpha}_j = \left(\frac{1}{\lambda_j}\right) \hat{\beta}_j \text{ for } j = 1, \dots, p .$$

- b) (5 points) Do the relative numerical sizes of the features matter in linear regression? (By the way, there are applications where some features can be much larger than others. Consider, for instance, regressing housing price on the number of rooms and the total square footage. For some statistical learning methods, care must be taken to ensure that all features are on the same scale.)

Problem 3: Forecasting Hyundai Elantra Sales (Adapted from Bertsimas 22.1) (70 points)

Nearly all companies seek to accurately predict future sales of their product(s). If the company can accurately predict sales before producing the product, then they can better match production with customer demand, thus reducing unnecessary inventory costs while being able to satisfy demand for their product.

In this exercise, you are asked to predict the monthly sales in the United States of the Hyundai Elantra automobile. 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 in the United States. We will use linear regression to predict monthly sales of the Elantra using economic indicators of the United States as well as (normalized) Google search query volumes. The data for this problem is contained in the file **Elantra142-Fall2018.csv**. Each observation in the file is for a single month, from January 2010 through July 2018. The variables are described in Table 1.

- a) (25 points) Start by splitting the data into a training set and testing set. The training set should contain all observations for 2010 through 2015. The testing set should have all observations for 2016, 2017, and 2018.

Consider just the four independent variables **Unemployment**, **ElantraQueries**, **CPI.Energy**, and **CPI.All**. Using your regression skills, choose a subset of these four variables and construct a regression model to predict monthly Elantra sales (**ElantraSales**). Try to choose which of the four variables to use in your model in order to build a high-quality linear regression model. Use the training set to build your model, and do not add any additional variables beyond the four indicated independent variables. Write a brief explanation (no more than one page, preferably less) – targeted to a statistically literate manager – describing how you decided on the variables to use in the model and the quality of the linear regression model’s predictions, as evaluated using the training set (there is no need to consider the test set for this part of the problem). Be sure to address the following in your explanation:

- i) What is the linear regression equation produced by your model, and how should one interpret the coefficients for the independent variables? Consider readability issues when writing down the equation.
 - ii) How did you select the variables to include in your linear regression model? Does it make sense to remove any variables due to multicollinearity among the four independent variables?
 - iii) Which of these variables are significant?
 - iv) Do the signs of the model’s coefficients make sense? Are you reasonably sure that the signs are correct?
 - v) How well does the model predict training set observations? Can you justify the model’s performance on the training data with a quantifiable metric?
- b) (15 points) Let us now try to further improve the linear regression model by modeling seasonality. In predicting demand and sales, seasonality is often very important since demand

Table 1: Variables in the dataset `Elantra142-Fall2018.csv`.

Variable	Description
MonthNumeric	The observation month given as a numerical value (1 = January, 2 = February, 3 = March, etc.).
MonthFactor	The observation month given as the name of the month (which will be a factor variable in R).
Year	The observation year.
ElantraSales	The number of units of the Hyundai Elantra sold in the United States in the given month and year.
Unemployment	The estimated unemployment rate (given as a percentage) in the United States in the given month and year.
ElantraQueries	A (normalized) approximation of the number of Google searches for “hyundai elantra” in the United States in the given month and year.
CPI.All	The consumer price index (CPI) for all products for the given month and year. This is a measure of the magnitude of the prices paid by consumer households for goods and services.
CPI.Energy	The monthly consumer price index (CPI) for the energy sector of the US economy for the given month and year.

for most products tends to be periodic in time. For example, demand for heavy jackets and coats tends to be higher in the winter, while demand for sunscreen tends to be higher in the summer.

Construct a new linear regression model using the **MonthFactor** variable as an independent variable, in addition to all four of the variables you used at the start of part (a). As before, construct your model based on the training data.

*Note: When **MonthFactor** is used in the linear regression model, you should see in the summary output that R created a regression coefficient for each of the possible values of the variable, except for one. R will always do this when a factor variable is used in a regression model; it will turn the variable into a bunch of categorical variables, each of which will have data values 0 or 1 in each of the observation records. For example, the variable **MonthFactorAugust***

should have the data value 1 if the observation month is August, and value 0 otherwise. The coefficient for this categorical variable will then get added to the regression prediction if and only if the observation month is August. (The reason why one of the factor levels is not listed is due to a redundancy. You can think of the impact of that level as being included in the intercept term.)

Answer the following questions about this modeling exercise.

- i) Describe your new model. What is the regression equation? (Do not simply copy and paste output from R.) How should one interpret the coefficients of each of the **MonthFactor** dummy variables?
 - ii) What is the training set R^2 for the new model? Which variables are significant?
 - iii) Do you think adding the independent variable **MonthFactor** improves the quality of the model? Why or why not?
 - iv) Can you think of a different way that you might use the given data to model seasonality? Do you think your new way would improve on the best model you have constructed so far? (By the way, later in the course we will have a lecture dedicated to basic time series modeling, and we will explore a number of ways to construct models using datasets with an associated time component.)
- c) (12 points) Build a final model using a subset of the independent variables used in parts (a) and (b), providing a brief justification for the variables selected. What is the training set R^2 and the OSR^2 (this is the R^2 of your model on the test set)? Do you think your model would be useful to Hyundai? Why or why not?
- d) (13 points) Let us now consider building a model for predicting the sales of a different car model. First, you will need to collect some new data. In particular, please do the following:
- i) There are various websites that publish U.S. car sales data. Navigate to one such site, <http://www.goodcarbadcar.net/category/sales-by-model/>, and select **one** of the following three car models whose sales you are interested in predicting:
 - Toyota Prius, data located at:
<http://www.goodcarbadcar.net/2011/01/toyota-prius-sales-figures/>
 - GMC Sierra, data located at:
<http://www.goodcarbadcar.net/2011/01/gmc-sierra-sales-figures/>
 - Honda Odyssey, data located at:
<http://www.goodcarbadcar.net/2011/01/honda-odyssey-sales-figures/>

Throughout, let us use “**zzz**” as a placeholder to refer to the name of the car model that you selected. Collect the monthly sales data for the model you selected and append it as a new column of the data file called **zzzSales**.
 - ii) Now navigate to the Google Trends site, <https://www.google.com/trends/explore>, and enter the search term corresponding to the full name of your selected car mode (either “toyota prius”, “gmc sierra”, or “honda odyssey”). Be sure that the region is set to the U.S. and that the time frame covers January 2010 through July 2018. Download the relevant queries data for your search term and append a new column to the data file called **zzzQueries**.

As in part (c), build a model to predict monthly sales of your selected car model, **zzzSales**, based on a subset of the independent variables used in parts (a) and (b) but with **ElantraQueries** replaced with **zzzQueries**. Be sure that **ElantraQueries** and **ElantraSales** are NOT used as independent variables. Provide a brief justification for the variables selected. What is the training set R^2 and the OSR^2 ? How does the performance of the model for predicting **ElantraSales** in part (c) compare to the performance of your new model for predicting **zzzSales**?

- e) (5 points) Using the models from parts (c) and (d), make predictions for August 2018 U.S. sales for both the Hyundai Elantra model and your selected car model. You may either search for relevant feature values on the Internet or make intelligent guesses based on the data provided. How confident are you in your predictions? (OPTIONAL, just for fun and not for credit: check <http://www.goodcarbadcar.net/> and, if the actual sales for August 2018 have been recorded, report, for both models, the absolute difference between your prediction and the corresponding actual sales figure.)