INDENG 242: Applications in Data Analysis, Fall 2023

Homework Assignment #1

September 7, 2023

1 Problem 1: (15 points)

I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$.

a. Suppose that the true relationship between X and Y is linear, $Y = \beta_0 + \beta_1 X + \varepsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

b. Answer (a) using test rather than training RSS.

c. Suppose that the true relationship between X and Y is not linear, but we don' t know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

d. Answer (c) using test rather than training RSS.

2 Problem 2: (20 points)

Consider the multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon \ . \tag{1}$$

Suppose that we have collected data $(x_1, y_1), \ldots, (x_n, y_n)$, where each x_i is an observed feature vector of length p and each y_i is an observed scalar response. Let $\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$ denote the sample mean of the dependent variable and let $\bar{x}_j := \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ denote the sample mean of the j^{th} feature for $j = 1, \ldots, p$. Define *centered* versions of the dependent variable and the features by $W := Y - \bar{y}$ and $Z_j := X_j - \bar{x}_j$ for $j = 1, \ldots, p$. Similarly, define centered versions of the observed data by:

$$w_i := y_i - \bar{y} \ \text{ for } i = 1, \ldots, n \ ,$$

and

$$z_{ij} := x_{ij} - \bar{x}_j \quad \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, p \; .$$

Now consider a different multiple linear regression model based on the centered dependent variable and centered features:

$$W = \alpha_0 + \alpha_1 Z_1 + \ldots + \alpha_p Z_p + \epsilon \; . \tag{2}$$

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

Please answer the following:

a) (10 points) Let $\hat{\alpha}$ denote the regression coefficient estimates with respect to model (2), trained on the data $(z_1, w_1), \ldots, (z_n, w_n)$. Show that $\hat{\alpha}_0 = 0$. (Hence, the intercept is not needed in model (2).)

b) (5 points) Let $\hat{\beta} := (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ denote the regression coefficient estimates for model (1), trained on the data $(x_1, y_1), \ldots, (x_n, y_n)$. It turns out that $\hat{\alpha}_j = \hat{\beta}_j$ for $j = 1, \ldots, p$, i.e., the coefficient estimates for the features do not change after centering. Please provide a one or two sentence intuitive explanation for why this is the case. (You do not need to provide a formal mathematical justification of this fact.) Together with part *(a)*, this implies that $\hat{\beta}_1, \ldots, \hat{\beta}_p$ can be estimated without an intercept.

c) (5 points) Given a new p-dimensional feature vector x_{new}, show how to use the results of training model (2) to generate a corresponding prediction for the dependent variable Y.

3 Problem 3: Forecasting Chevrolet Silverado Sales (Adapted from Bertsimas 22.1) (65 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 Chevrolet Silverado automobile. We will use linear regression to predict monthly sales of the Silverado 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 **`Silverado242-Fall2023.csv`**. Each observation in the file is for a single month, from January 2010 through June 2020. 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 through 2020.

Table 1: Variables in the dataset `Silverado242-Fall2023.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.
`SilveradoSales`	The number of units of the Silverado 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.
`SilveradoQueries`	A (normalized) approximation of the number of Google searches for "silverado" 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.

Consider just the four independent variables `Unemployment`, `SilveradoQueries`, `CPI.Energy`, and `CPI.All`. Using your regression skills, choose a subset of these four variables and construct a regression model to predict monthly Silverado sales (`SilveradoSales`). 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 (e.g., do not just copy and paste the output from R).

ii) How did you select the variables to include in your linear regression model?

iii) Do the signs of the model's coefficients make sense? Are you reasonably sure that the signs are correct?

iv) 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 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)*. **There is no need to do variable selection for this part of the problem.** As before, construct your model based on the training data.

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) (15 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 Chevrolet? Why or why not?

d) (10 points) In regression analysis, a loss function ℓ takes as input a predicted value $\hat{y}$ and an observed value y, and it returns a value $\ell(\hat{y}, y)$ which is interpreted as the loss/error/cost associated with predicting $\hat{y}$ when the actual value of the dependent variable is y. So far, we have only considered the squared loss $\ell(\hat{y}, y) := (y - \hat{y})^2$, which is the most standard loss function in regression. However, the squared loss is not always the most appropriate or the most effective in every situation.

Consider the following (greatly simplified) scenario regarding how Chevrolet makes monthly production decisions. Firstly, the management at Chevrolet has decided to use the predictions of your regression model to directly set monthly inventory levels. That is, if your model predicts that next month's Silverado sales will be $\hat{y}$, then Chevrolet will have available exactly $\hat{y}$ Silverado units to be sold next month. (You may ignore integer constraint issues and assume that Chevrolet can produce fractional units.) Whenever a unit is not sold in a given month, then Chevrolet can use that unit to offset part of next month's production. For example, if Chevrolet has five Silverado units available in January but only sells three of them, then they can carryover two units to February. For simplicity, you may assume 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. Finally, there is a cost of \$500 associated with carrying over a unit from one month to the next. Suppose that Chevrolet earns a profit of \$3000 for each Silverado unit that it sells. Propose a loss function ℓ that accurately models this situation and explain your reasoning.