IEOR 242: Applications in Data Analysis, Fall 2019

Homework Assignment #1

September 30, 2019

Problem 1: (15 points)

Consider the multiple linear regression model:

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

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

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

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. Similarly as before, define the rescaled observed features by:

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

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 $\text{rank}(\mathbf{X}) = p + 1 < n$.

Please answer the following:

a) (10 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 $(x_1, y_1), \ldots, (x_n, y_n)$. Likewise, let $\hat{\alpha}$ denote the regression coefficient estimates with respect to model (2), trained on the data $(z_1, y_1), \ldots, (z_n, y_n)$. Show that $\hat{\alpha}_0 = \hat{\beta}_0$ and

$$\hat{\alpha}_j = (\tfrac{1}{\lambda_j})\hat{\beta}_j \ \text{ for } j = 1, \ldots, p \ .$$

Answer

In the following, let

$$RSS_1(\beta) = \|y - X\beta\|_2^2$$

$$RSS_2(\alpha) = \|y - Z\alpha\|_2^2$$

We provide two different methods to solve this problem:

Method 1: Since $\hat{\beta} = \arg\min_{\beta} RSS_1(\beta)$, then there is $2(y - X\hat{\beta})^T X = 0$, and by X full rank,

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Similarly $\hat{\alpha} = \arg\min_{\alpha} RSS_2(\alpha)$, we have

$$\hat{\alpha} = (Z^T Z)^{-1} Z^T y.$$

Note that $z_{ij} := \lambda_j x_{ij}, \forall i = 1, \ldots, n \forall j = 1, \ldots, p$ is equivalent to

$$Z = X\Lambda.$$

Where $\Lambda = Diag(1, \lambda_1, \ldots, \lambda_p)$ is a $(p+1) \times (p+1)$ diagonal matrix with $(1, \lambda_1, \ldots, \lambda_P)$ as elements in the diagonal. Plug in, we get

$$\begin{aligned}
\hat{\alpha} &= (Z^T Z)^{-1} Z^T y \\
&= ((X\Lambda)^T X\Lambda)^{-1} (X\Lambda)^T y \\
&= (\Lambda X^T X \Lambda)^{-1} \Lambda X^T y \\
&= \Lambda^{-1} (X^T X)^{-1} \Lambda^{-1} \Lambda X^T y \\
&= \Lambda^{-1} (X^T X)^{-1} X^T y \\
&= \Lambda^{-1} \hat{\beta}
\end{aligned}$$

Note that $\Lambda^{-1} = Diag(1, \lambda_1^{-1}, \ldots, \lambda_p^{-1})$, we have $\hat{\alpha}_0 = \hat{\beta}_0$ and $\hat{\alpha}_j = \lambda_j^{-1} \hat{\beta}_j$ for $j = 1, \ldots, p$.

Grading criteria: Max score for this question: **12 points**.

- **3 point**: Get $\hat{\beta} = (X^T X)^{-1} X^T y$.
- **3 points**: Get $\hat{\alpha} = (Z^T Z)^{-1} Z^T y$
- **3 points**: Get $Z = X\Lambda$ or anything equivalent to this.
- **3 points**: Get the final result using algebra.

Method 2: We define α^* with $\alpha_0^* := \beta_0^*$ and $\alpha_j^* := \frac{1}{\lambda_j} \hat{\beta}_j$ for $j = 1, \ldots, p$. And define β^* with $\beta_0^* := \hat{\alpha}_0$, $\beta_j^* := \lambda_j \hat{\alpha}_j$ for $j = 1, \ldots, p$. Note that $\hat{\beta}$ minimizes the RSS_1, and $\hat{\alpha} minimizes the RSS_2$.

We first show that $RSS_2(\alpha^*) = RSS_1(\hat{\beta})$:

$$
\begin{aligned}
\mathrm{RSS}_2(\alpha^*) &= \sum_{i=1}^{n}(y_i - \alpha_0^* - \sum_{j=1}^{p}\alpha_j^* z_{ij})^2 \\
&= \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\frac{\hat{\beta}_j}{\lambda_j} z_{ij}\right)^2 \\
&= \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\frac{\hat{\beta}_j}{\lambda_j}\lambda_j x_{ij}\right)^2 \\
&= \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2 \\
&= \mathrm{RSS}_1(\hat{\beta})
\end{aligned}
$$

Then we show that $\mathrm{RSS}_1(\beta^*) = \mathrm{RSS}_2(\hat{\alpha})$:

$$
\begin{aligned}
\mathrm{RSS}_1(\beta^*) &= \sum_{i=1}^{n}(y_i - \beta_0^* - \sum_{j=1}^{p}\beta_j^* x_{ij})^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{\alpha}_0 - \sum_{j=1}^{p}\hat{\alpha}_j\lambda_j x_{ij})^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{\alpha}_0 - \sum_{j=1}^{p}\hat{\alpha}_j\lambda_j \frac{z_{ij}}{\lambda_j})^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{\alpha}_0 - \sum_{j=1}^{p}\hat{\alpha}_j z_{ij})^2 \\
&= \mathrm{RSS}_2(\hat{\alpha})
\end{aligned}
$$

Note that $\hat{\beta}$ minimize the RSS_1. Hence $RSS_1(\hat{\beta}) \leq RSS_1(\beta^*)$. Using the same argument, we have $RSS_2(\hat{\alpha}) \leq RSS_2(\alpha^*)$. Together with $RSS_1(\hat{\beta}) \leq RSS_1(\beta^*)$ and $RSS_2(\hat{\alpha}) \leq RSS_2(\alpha^*)$, we have

$$
RSS_1(\hat{\beta}) = RSS_2(\hat{\alpha})
$$

$$
\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 = \sum_{i=1}^{n}(y_i - \hat{\alpha}_0 - \sum_{j=1}^{p}\hat{\alpha}_j\lambda_j x_{ij})^2
$$

We have a single formula for $\hat{\beta}$ according to slides, indeed, people can prove it is unique. Hence we can conclude that $\hat{\alpha}_0 = \hat{\beta}_0$ and $\hat{\alpha}_j\lambda_j = \hat{\beta}_j$.

Grading criteria: Max score for this question: **12 points**.

- **3 point**: Get $\mathrm{RSS}_2(\alpha^*) = \mathrm{RSS}_1(\hat{\beta})$.
- **3 points**: Get $\mathrm{RSS}_1(\beta^*) = \mathrm{RSS}_2(\hat{\alpha})$.
- **3 points**: Get $RSS_1(\hat{\beta}) = RSS_2(\hat{\alpha})$.

- **3 points**: Get the final result (The final result is stronger than $RSS_1(\hat{\beta}) = RSS_2(\hat{\alpha})$, can not directly get from this equation).

b) (3 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.)

Answer
The relative numerical sizes of the features **does not matter** in linear regression. This is because changes to the scale of the features are captured in the values of the coefficients. This is what we have proven in (a).

Problem 2: (20 points)

Again, consider the multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon \ . \tag{3}$$

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

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

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

$$W = \alpha_0 + \alpha_1 Z_1 + \ldots + \alpha_p Z_p + \epsilon \ . \tag{4}$$

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 (4), 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 (4).)
Answer 1 Let Z denote the $n \times p$ matrix without the augmented column. For any W, Z, the OLS estimators $\hat{\alpha}_0$ and $\hat{\alpha} = (\hat{\alpha}_1, ..., \hat{\alpha}_p)$ are

$$\hat{\alpha}_0, \hat{\alpha} = \text{argmin}_{\hat{\alpha}_0, \hat{\alpha}} \|W - Z\alpha - \alpha_0\|^2 \tag{5}$$

Let $RSS(\alpha_0, \hat{\alpha}) = \|W - Z\alpha - \alpha_0\|^2$. The partial derivative of α_0 is

$$\frac{\partial RSS}{\partial \alpha_0} = -2\sum_{i=1}^{n}(w_i - z_i^T\alpha - \alpha_0) \tag{6}$$

According to the first-order condition, $\hat{\alpha}_0$ satisfies

$$-2\sum_{i=1}^{n}(w_i - z_i^T\hat{\alpha} - \hat{\alpha}_0) = 0 \tag{7}$$

$$\Leftrightarrow \hat{\alpha}_0 = \frac{\sum_{i=1}^{n}(y_i - x_i^T\hat{\alpha})}{n} = \bar{W} - \bar{Z}\hat{\alpha} \tag{8}$$

When $W = Y - \bar{Y}$ and $Z_j = X_j - \bar{X_j}$ for all $j = 1, ..., p$, $\bar{W} = 0$ and $\bar{Z} = 0$. Then by (8), $\hat{\alpha}_0 = \bar{W} - \bar{Z}\hat{\alpha} = 0$

Grading criteria: Max score for this question: **10 points**.

- **3 point**: Get the first-order condition.
- **4 points**: Get the formula for the intercept.
- **2 points**: Get the mean of W and Z is zero.
- **1 points**: Get the final conclusion.

Answer 2

The linear equation system of the first order condition can be written as:

$$Z^T Z\hat{\alpha} = Z^T W$$

By matrix multiplication, one can verify that the first row of $Z^T Z$ is

$$[n, \sum_{i=1}^{n} Z_{i1}, \sum_{i=1}^{n} Z_{i2}, ..., \sum_{i=1}^{n} Z_{ip}]$$

Similarly, the first element of $Z^T W$ is $\sum_{i=1}^{n} W_i$

Recall that $W = Y - \bar{Y}$ and $Z_j = X_j - \bar{X_j}$ for all $j = 1, ..., p$. Therefore, $\sum_{i=1}^{n} Z_{ij} = 0$, $\forall j = 1, ..., p$. $\sum_{i=1}^{n} W_i = 0$. The first linear equation in the linear equation system is $n\hat{\alpha}_0 = 0$.

Therefore, $\hat{\alpha}_0 = 0$.

Grading criteria: Max score for this question: **10 points**.

- **3 point**: Get the first-order condition.
- **4 points**: Get the matrix multiplication results.
- **2 points**: Get the mean of W and Z is zero.
- **1 points**: Get the final conclusion.

b) (5 points) Let $\hat{\beta} := (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ denote the regression coefficient estimates for model (3), 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.
Answer $(\hat{\beta}_1, ..., \hat{\beta}_p)$ stands for the slope of the regression function. Centering the independent and dependent variables is just shifting their positions in the $\mathbb{R}^{p+1}$ space. Therefore the resulting slope should be the same before and after centering.

c) (5 points) Given a new p-dimensional feature vector x_{new}, explain how to use the results of training model (4) to generate a corresponding prediction for the dependent variable Y.
Answer

$$\hat{y}_{new} = \hat{\beta}^T x_{new} + \hat{\beta}_0 \tag{9}$$
$$= \hat{\alpha}^T (x_{new} - \bar{X}) + \bar{Y} \tag{10}$$

Problem 3: Forecasting Jeep Wrangler 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 Jeep Wrangler automobile. Jeep is brand of American automobiles that is a division of the Italian-American corporation Fiat Chrysler Automobiles (FCA). The Wrangler is a car model that has been produced since 1986, with most of its sales in the United States. We will use linear regression to predict monthly sales of the Wrangler 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 `Wrangler242-Fall2019.csv`. Each observation in the file is for a single month, from January 2010 through June 2019. 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 2019.

Consider just the four independent variables `Unemployment`, `WranglerQueries`, `CPI.Energy`, and `CPI.All`. Using your regression skills, choose a subset of these four variables and construct a regression model to predict monthly Wrangler sales (`WranglerSales`). 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

Table 1: Variables in the dataset `Wrangler242-Fall2019.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.
`WranglerSales`	The number of units of the Jeep Wrangler 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.
`WranglerQueries`	A (normalized) approximation of the number of Google searches for "jeep wrangler" 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.

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?

Answer

There is more than one way of generating the final model. In our case, we went in the following order:

- `Unemployment`, `WranglerQueries`, `CPI.Energy`, and `CPI.All`
  Remove `Unemployment` due to multicollinearity with CPI variables (large VIF) together with domain-specific information (we expect unemployment to influence car sales less than CPI index, although that is debatible).
- `WranglerQueries`, `CPI.Energy`, and `CPI.All`
  Remove `CPI.All` because it was not significant up to a 95% confidence level with large VIF. However, this is debatible as well (it is also true with WranglerQueries, but we consider the latter to be more correlated to the sales). $R^2 \approx 79\%$.
- `WranglerQueries`, `CPI.Energy`
  All VIF values are OK but still CPI.Energy is not significant ($R^2 \approx 79\%$).
- `WranglerQueries`
  This is our final model. All of our variables are highly significant and we have only a slight deteriorition in goodness-of-fit to training data ($R^2 \approx 79\%$).

Grading crtieria. Max score for this question: **25 points**.

- **4 points**: Clarity of response
- **5 points**: Explicitly showing final regression model. Including representative names for the variables together with their coefficients (not just R output).
- **6 points**: General guidance to how final model was reached, indicating whether variables were discarded because their variability was not directly related to response variable (i.e. low p-value) or because the information captured by that variable is already contained in the other independent variables (i.e. multicollinearity).
- **5 points**: Interpretation of the model. Do coefficients signs make sense? etc.
- **5 points**: How well does the model perform on training data? Make sure a mention to a quantifiable metric is used.

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.

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. As discussed in class, 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.)

Answer Here, the model with four variables together with the `MonthFactor` should be used (from problem statement).

Grading crtieria. Max score for this question: **13 points**.

- **2 points**: Regression equation clearly shown (not just R output).
- **3 points**: Interpretation of `MonthFactor` dummy variables: extra sales in a given month compared to the baseline month keeping other variables constant/fixed.
- **2 points**: R^2 of new model together with analysis of which variables are significant. We should expect a significant increase in R^2 (I went from around 79% to 87% aprox).
- **2 points**: Conclusion that `MonthFactor` improves model quality.
- **2 points**: Explanation of why this is the case: we get a significantly better fit without multicollinearity.
- **2 point**: Reasonable alternative to model seasonality (such as model according to fall, summer, spring and winter given that many of the dummy variables are not statistically significant).

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 Jeep/FCA? Why or why not?

Answer There is no unique correct model here, and the justifications will tend to repeat what is shown in part (a) or (b). (I use `CPI.Energy` , `WranglerQueries` and `MonthFactor` to construct my final model and training set $R^2 = 0.87$ and $OSR^2 = 0.62$).

Grading criteria. Max score for this question: **12 points**.

- **5 points**: For choosing a final model which is reasonable (**2 points**), providing a brief explanation of why that model was chosen (it can be either at the individual variable level or referring back to arguments in parts (a) and (b)) (**3 points**).
- **5 points**: For providing an R^2 and OSR^2 that is in the ballpark of what to be expected.
- **5 points**: For providing an reasonable explanation as to why model is useful or not (for example: $OSR^2 > 0$ means that we improve compared to the baseline model: hence the model is useful).

d) (10 points) Let us now consider adding an additional feature/variable to your final model from part *(c)*. Based on your knowledge and intuition, think of a monthly variable that you hypothesize might be related to Wrangler sales. Provide a one or two sentence explanation for your choice. Search online for a data source for your chosen variable (if you are not able to find data, then you need to pick a different variable), and append your collected data as a new column in the original data file. (It is OK to use variables similar to what we used above, i.e., a different economic indicator or Google trends data for a different search term, but feel free to get as creative as you like.)

Now, build a new regression model with your additional chosen feature in addition to the features that you selected in part *(c)*. Does the new feature add any predictive value? Justify your answer based on the results of your analysis.

Answer This is an open ended question. Grading scale (note that answers fall under only under one category here). Max score for this question: **10 points**.

- **5 points** for getting the data.
- **5 points** for building and interpretating the model.
  - A new feature with a reasonable metric gets full credit: **5 points**.
  - Features are provided without explanation: **2 points**.