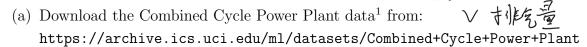
- 1. Prove the Gauss-Markov Theorem, i.e. show that the least squares estimate in linear regression is the BLUE (Best Linear Unbiased Estimate), which means  $Var(\mathbf{a}^T \widehat{\beta}) \leq Var(\mathbf{c}^T \mathbf{y})$  where  $\mathbf{c}^T \mathbf{y}$  is any unbiased estimator for  $\mathbf{a}^T \beta$ . (20 pts)
- 2. (**Linear Regression with Orthogonal Design**) Assume that the columns  $\mathbf{x}_0, \dots, \mathbf{x}_p$  of  $\mathbf{X}$  are orthogonal. Express  $\widehat{\beta}_i$  in terms of  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_p$  and  $\mathbf{y}$ . (10 pts)
- 3. (The Minimum Norm Solution) When  $\mathbf{X}^T\mathbf{X}$  is not invertible, the normal equations  $\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{y}$  do not have a unique solution. Assume that  $\mathbf{X} \in \mathbb{R}_r^{n \times (p+1)}$ , where r is the rank of  $\mathbf{X}$ . Assume that the SVD of  $\mathbf{X}$  is  $\mathbf{U}\Sigma\mathbf{V}^T$ , where  $\mathbf{U} \in \mathbb{R}^{n \times r}$  satisfies  $\mathbf{U}^T\mathbf{U} = \mathbf{I}_r$ . Also  $\mathbf{V} \in \mathbb{R}^{(p+1)\times r}$  satisfies  $\mathbf{V}^T\mathbf{V} = \mathbf{I}_r$  and  $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \dots, \sigma_r)$  is the diagonal matrix of positive singular values.
  - (a) Show that  $\beta_{\text{mns}} = \mathbf{V} \mathbf{\Sigma}^{-1} \mathbf{U}^T \mathbf{y}$  is a solution to the normal equations. (5 pts)
  - (b) Show that for any other solution  $\beta$  to the normal equations,  $\|\beta\| \ge \|\beta_{\text{mns}}\|$ . [Hint: one way (and not the only way) of doing this is to show that  $\beta = \beta_{\text{mns}} + b$ .] (15 pts)
  - (c) Is  $\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T$  the pseudo-inverse of  $\mathbf{X}$ ? (Hint: you can prove or disprove using the so-called Penrose properties) (10 pts)



- (b) Exploring the data: (5 pts)
  - i. How many rows are in this data set? How many columns? What do the rows and columns represent?
  - ii. Make pairwise scatterplots (scatter matrix) of all the variables in the data set including the predictors (independent variables) with the dependent variable. Describe your findings.
  - iii. What are the mean, the median, range, first and third quartiles, and interquartile ranges of each of the variables in the dataset? Summarize them in a table.
- (c) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back



<sup>&</sup>lt;sup>1</sup>There are five sheets in the data. All of them are shuffled versions of the same dataset. Work with Sheet 1.

- up your assertions. Are there any outliers that you would like to remove from your data for each of these regression tasks? (10 pts)
- (d) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis  $H_0: \beta_j = 0$ ? (10 pts)
- (e) How do your results from 4c compare to your results from 4d? Create a plot displaying the univariate regression coefficients from 4c on the x-axis, and the multiple regression coefficients from 4d on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis. (5 pts)
- (f) Is there evidence of nonlinear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form<sup>2</sup>

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

- (g) Is there evidence of association of interactions of predictors with the response? To answer this question, run a full linear regression model with all pairwise interaction terms and state whether any interaction terms are statistically significant. (5 pts)
- (h) Can you improve your model using possible interaction terms or nonlinear associations between the predictors and response? Train the regression model on a randomly selected 70% subset of the data with all predictors. Also, run a regression model involving all possible interaction terms  $X_iX_j$  as well as quadratic nonlinearities  $X_j^2$ , and remove insignificant variables using p-values (be careful about interaction terms). Test both models on the remaining points and report your train and test MSEs. (10 pts)
- (i) KNN Regression:
  - i. Perform k-nearest neighbor regression for this dataset using both normalized and raw features. Find the value of  $k \in \{1, 2, ..., 100\}$  that gives you the best fit. Plot the train and test errors in terms of 1/k. (10 pts)
- (j) Compare the results of KNN Regression with the linear regression model that has the smallest test error and provide your analysis. (5 pts)

<sup>&</sup>lt;sup>2</sup>https://scikit-learn.org/stable/modules/preprocessing.htm\#generating-polynomial-features