STAT 444/844 , CM 764
Assignment #1 - (Spring 2020)

- **Due:** Tuesday June 2 at 11:59pm
- **Submission platform:** You must upload your solutions in the form of **one pdf file for each question** by the deadline onto Learn (a total of 5 pdf files for undergrads and 6 pdf files for grads).Your instructor will NOT accommodate mistakes in submitting the pdf file of one question for another question. No assignment submission through email will be accepted. You are only allowed to submit one file for each question and cannot change your file once submitted.
- **Format:** Your assignment must be generated using LaTex or RMarkdown, with no hand-written portions and/or screenshots in your submitted files. Taking a picture of what you would like to submit and embedding it as an image within a LaTeX or RMarkdown file counts as submitting hand-written solution and/or a screenshot which will be ignored for marking purposes. The individual parts of each question (a, b, c, ...) must start on top of a new page. Notice that the presentation of your paper, i.e. format, clarity, and providing graphs and codes **integrated within the solution** usually affects marking. A clear and well-organized paper is easier to mark. Your paper may not be graded or may receive significant mark deduction penalty if you do not follow the required format outlined here.
- **Late submissions:** The deadline is firm and late submissions and/or questions with no submitted files will automatically receive a grade of 0%. Given the high weight of the assignment, I strongly suggest that you set a soft deadline of an earlier time to submit the assignment so that you will not be penalized for technical issues which usually happen during the last-minute submissions.
- **Grading scheme:** For theoretical questions, make sure that you provide the detailed steps of the derivation/proof. For questions where R is used, both the codes and the R outputs along with your interpretations and answers to the questions must be submitted. Your R codes and outputs must be integrated within the solutions to each part of the question and must NOT be put in an appendix and/or at the end of the question.
- **Academic integrity:**
  - You may not talk to any other individual about the questions in this assignment. The instructor will hold online office hours during which he will answer clarification questions.
  - You may not use and/or search the internet (except for Learn and Piazza) to answer the questions in this assignment.
  - In short, you can treat this assignment like an open-book exam, where you are only allowed to use the course material provided to you during lectures and on Piazza and/or Learn as well as the books introduced in the course ouline.

  Any violation of the the academic integrity regulations outlined here and in the course outline will be counted as cheating and will be reported to the Dean's Office. I remind you that Turnitin is used for this assignment.

  You understand that the instructor reserves the right to conduct an online interview with you during which you will be asked questions about you solutions and the details of how you came to these responses. Should such an interview take place and you are unable to explain and defend your solutions, your grade for this assignment, and consequently, your course grade will be affected.

1. Consider the regression model $Y_i = \mu(\mathbf{x}_i) + \epsilon_i$ in which $\mu(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ and $\epsilon_i \overset{\text{ind.}}{\sim} N(0, \sigma^2 g(x_i))$.

   (a) **[5 marks]** Show that the weighted least squares estimator $\widetilde{\boldsymbol{\beta}}_{WLS}$ is an unbiased estimator of $\boldsymbol{\beta}$ where W is the the diagonal matrix of the weights $w_1, ..., w_n$.

   (b) **[3 marks]** Under what choice of the weight matrix W do you we get $Var(\widetilde{\boldsymbol{\beta}}_{WLS}) = \sigma^2(X^T W X)^{-1}$? Note that if Z is a random vector and A is a constant matrix, then $Var(AZ) = A \times Var(Z) \times A^T$.

The manufacturing company JAX produces a particular type of aluminum screw which is used in the aviation industry. The data-set `JaxSales.txt` includes the annual number screws sold by 15 distributes of this company over a period of 20 years: 1991-2010. We would like to model Sales (Y) as a function of Year (X) using a polynomial regression model of degree p, where $p \in \{1, 2, ..., 10\}$. In questions 2-4 you will be working on the dataset `JaxSales.txt`.

Note: Fitting polynomial regression models is easier and more stable using the function `poly(x,q)` which returns orthogonal polynomials of degree 1 to q over the set of points x. For example, to fit a regression model of degree 4, you can write `lm(y ~ poly(x,4))`.

2. Considering the model $Y_i | X_i = \beta_0 + \beta_1 X_i + ... + \beta_p X_i^p + \epsilon_i$, where $\epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$, we would like to choose p using cross-validation.

   (a) Use LOO cross-validation to choose the polynomial degree p. You should:

      i. **[4 marks]** Provide the plot of Cross-Validation MSE vs. p, $p = 1, 2, ..., 10$.

      ii. **[3 marks]** Generate the scatterplot of the data and superimpose the fitted model chosen by LOOCV.

   (b) Use 10-fold cross-validation to choose the polynomial degree p. You should:

      i. **[4 marks]** Provide the plot of Cross-Validation MSE vs. p, $p = 1, 2, ..., 10$.

      ii. **[3 marks]** Generate the scatterplot of the data and superimpose the fitted model chosen by 10-fold cross-validation.

   (c) **[3 marks]** Comparing (a) and (b), do the two methods result in different models? Which of the two values of k do you prefer here? Why?

---

3. Reading the scatter plot of Sales (y) vs. Year (x) in the previous question, it is clear that the variability in y is an increasing function of x. We propose the model $Y_i | X_i = \beta_0 + \beta_1 X_i + ... + \beta_p X_i^p + \epsilon_i$, where $\epsilon_i \overset{\text{ind.}}{\sim} N\left(0, \sigma^2(x)\right)$ where $\sigma^2(x) = \alpha_0 + \alpha_1 x$.

   (a) **[4 marks]** Define the vector `Year.x` with values 1991, ..., 2010. For each element of `Year.x`= 1991, 1992, ..., 2010, calculate the variance of `Sales` for the corresponding year in `Year.x`. Save these variance values in the vector `sigma2.y`, and plot `sigma2.y` versus `Year.x`.

   (b) **[3 marks]** Fit a simple linear regression model of the form $sigma2.y = \alpha_0 + \alpha_1 Year.x$. Generate the scatter plot of `sigma2.y` versus $Year.x$ and add the fitted line to the plot. Report the values of $\widehat{\alpha}_0, \widehat{\alpha}_1$.

(c) **[5 marks]** Fit a weighted least squares polynomial regression model with degree chosen in question 2 and two choices of weight function:

1) $w_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 x_i$ and
2) $w_i = 1/(\widehat{\alpha}_0 + \widehat{\alpha}_1 x_i)$.

For each choice of weights, plot the data, along with the fitted model. Use the `predict` function in R, and add 95% prediction intervals to the plot.
Note: When using the predict function to produce prediction intervals for models fitted by weighted least-squares, the weights must be specified as an argument as well as the fitted model.

(d) **[5 marks]** Compare the plots in part (c) and (d) and comment on your finding. For each of the weight functions, which points in the scatterplot of Sales versus Year would have the greatest influence on the fitted line? Which would have the least? Which weight function would you choose and why?

---

4. Another way to calculated the weights for the weighted least squares is to estimate the the variance $\sigma(x)$ at each value of x based on the repeated measurements. Consider the weights $w_i = 1/\widehat{\sigma}^2(x_i)$ where $\widehat{\sigma}^2(x_i)$ is the variance of `Sales` for the year x_i.

(a) **[3 marks]** Fit a weighted least squares polynomial regression model with degree chosen in question 2 and the weight function explained above. Plot the data, along with the fitted model. Use the `predict` function in R, and add 95% prediction intervals to the plot.

(b) **[5 marks]** Provide the standard errors of the estimated parameters of

- the model chosen in question 2,
- the two models in question 3(c),
- the model in question 4(a).

What do you observe? Which model do you choose and why?

---

5. **[5 marks]** From the material discussed on model selection and brass-variance trade-off, construct one true/false question and explain why it is true or false.
Rubric: Concept and difficulty, clarity, creativity [3 marks] – Explanation [2 marks]

---

6. **Graduate Students Only:** Consider the model $\mathbf{Y}|X = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with p explanatory variates where the vector $\boldsymbol{\epsilon}_{n\times 1}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix Σ. We fit a generalized least squares model with the weight matrix W.

(a) **[6 marks]** Derive the distribution of $\widetilde{\boldsymbol{\beta}}_{GLS}$ assuming $W = c\Sigma^{-1}$ for some constant Σ. You must provide the details of your derivations.

(b) **[4 marks]** Write the likelihood function of $\boldsymbol{\beta}$ and show that $\widetilde{\boldsymbol{\beta}}_{ML} = \widetilde{\boldsymbol{\beta}}_{GLS}$ if $W = c\Sigma^{-1}$ for some constant c.