Final Exam
May 17, 2019

STA3010 Regression Analysis

Prof. Feng YIN

Last name: _____

First name: _____

Student ID: _____

Major: _____

1. 1st attempt: ☐ 2. 2nd attempt: ☐ 3. 3rd attempt: ☐

Email: _____

Question	Q1	Q2	Q3	Q4	Bonus	Total	Grade
Mark	25	35	25	15	10	110	

Question 1: Generalized Least-Squares Applied to Signal Processing (25ps in total)

We focus on a multiple linear regression model given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{1}$$

where

- $\beta_j, j = 0, 1, 2, ..., k$ are the unknown model parameters;
- $x_j, j = 1, 2, ..., k$ are the inputs, deterministic and precisely known;
- ε is the random error;
- y is the output.

Given a dataset with n data points, the above multiple linear regression model can be written in a compact matrix form as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where

- $\mathbf{y} = [y_1, y_2, ..., y_n]^T$ is an $n \times 1$ vector of the outputs;
- $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_k]^T$ is a $p \times 1$ vector of the unknown model parameters with $p = k + 1$ and $p \ll n$;
- $\mathbf{X}$ is an $n \times p$ design matrix of the inputs, which is of full rank p;
- $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]^T$ is an $n \times 1$ vector of random error terms.

Note: We assume the random error terms follow a joint Gaussian distribution $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. **Both the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are assumed to be precisely known. Moreover, the covariance matrix is assumed to be positive definite (PD).**

Please solve the following questions:

1. (6 points) Please derive an unbiased generalized least-squares (GLS) estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_{GLS}$.
2. (6 points) Please derive the covariance matrix of the above derived GLS estimator, $\hat{\boldsymbol{\beta}}_{GLS}$.

Next, we want to apply the above model to a famous signal propagation modeling problem. Concretely, we want to study a wireless propagation environment by letting a wireless transmitter broadcast signals to a number of N receivers. The received signal strength on each receiver is denoted y_i, $i = 1, 2, ..., N$, and satisfy the following signal model

$$y_i = \beta_0 e^{\beta_1 x_i} \varepsilon_i, \tag{3}$$

where

- x_i is the known input, measuring the distance between the transmitter and the i-th receiver;
- y_i is the output, measuring the signal strength received by the i-th receiver;
- β_0 is a **known** path-loss at a reference distance (for instance at 1 meter); while β_1 is an **unknown** path-loss exponent;
- ε_i is a multiplicative random error instead of additive random error.

Various studies have disclosed that the random error ε_i follows log-normal distributions, in other words, $\ln(\varepsilon_i) \sim \mathcal{N}(0, \sigma^2(x_i))$. Herein, $\sigma^2(x_i)$ means that the noise variance for $\ln(\varepsilon_i)$ is a known function of x_i. Concretely, we let $\sigma^2(x_i) = \frac{x_i^2}{10}$. We assume that the random error terms are independent.

Please solve the following problems:

1. (8 points) With the aid the above results, please give the final expression of the generalized leaset-squares (GLS) estimator of β_1 and give out the concrete expressions for the output vector $\boldsymbol{y}$, design matrix $\boldsymbol{X}$, and the "weight matrix" $\boldsymbol{\Sigma}$, etc.
2. (2 points) Do you think the derived GLS estimator of β_1 is Gaussian distributed? Why?
3. (3 points) Suppose in the signal propagation model in Eq. (3), there is an extra additive random error term $\tilde{\epsilon}_i$ which follows a zero-mean Gaussian distribution and uncorrelated with the multiplicative random error term ε_i, what kind of influence of would it bring to our GLS solution. Sketch your idea briefly. **Open ended.**

Question 2: Influential Points, Outliers, and Robust Regression (35ps in total)

We focus on a multiple linear regression model given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{4}$$

where

- $\beta_j, j = 0, 1, 2, ..., k$ are the unknown model parameters.
- $x_j, j = 1, 2, ..., k$ are the inputs. They are deterministic and precisely known.
- ε is the random error.
- y is the output.

Given a dataset with n data points, the above multiple linear regression model can be written in a compact matrix form as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5}$$

where

- $\mathbf{y} = [y_1, y_2, ..., y_n]^T$ is an $n \times 1$ vector of the outputs.
- $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_k]^T$ is a $p \times 1$ vector of the unknown model parameters. Herein, we let $p = k + 1$ for brevity and further assume $p \ll n$.
- $\mathbf{X}$ is an $n \times p$ design matrix of the inputs, which is of full rank p. Concretely,

$$\mathbf{X} = \begin{bmatrix} 1, & x_{1,1}, & x_{1,2}, & ..., & x_{1,k} \\ 1, & x_{2,1}, & x_{2,2}, & ..., & x_{2,k} \\ \vdots, & \vdots, & \vdots, & \ddots, & \vdots \\ 1, & x_{n,1}, & x_{n,2}, & ..., & x_{n,k} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \tag{6}$$

  where $\mathbf{x}_i^T$ represents the i-th row of the design matrix $\mathbf{X}$. Moreover, $\mathbf{X}^T\mathbf{X}$ is assumed to be nonsingular.

- $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]^T$ is an $n \times 1$ vector of random error terms.

Important observation of a given dataset: Seemingly, the majority of the outputs are contaminated by zero mean Gaussian i.i.d. random error terms. But there also exist a small fraction of outliers. Herein, the outliers are data points with ordinary input $\mathbf{x}$ values but abnormal output y values.

Please solve the following problems:

1. (2 points) For this problem, the hat matrix $\mathbf{H}$ is defined to be $\mathbf{H} \triangleq \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, please prove that the hat matrix $\mathbf{H}$ is idempotent.

2. (3 points) Every data point has a leverage score which is the i-th diagonal entry of the hat matrix $\mathbf{H}$. Can you show that the average leverage score over the n data points, i.e., $\frac{1}{n}\sum_{i=1}^{n} h_{ii}$, is equal to $\frac{p}{n}$, where p is number of columns and n is number of rows of the design matrix, $\mathbf{X}$. **Hint: H is an idempotent matrix.**

3. (3 points) You are given the Cook's influence measure D_i for each data point, taking the i-th data point for instance, as follows:

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)^T \mathbf{X}^T\mathbf{X} \left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)}{p \cdot MS_{Res}}. \tag{7}$$

   Please explain (in simple words) the meanings of $\hat{\boldsymbol{\beta}}_{(i)}$, $\hat{\boldsymbol{\beta}}$ and MS_{Res} in equation (7).

4. (6 points) Combine Eq.(7) given in the third sub-problem and the following hint

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_i e_i}{1 - h_{ii}} \tag{8}$$

   where e_i is the i-th ordinary residual obtained from the least-squares fit to all n data points; Eventually derive the Cook's influence measure for the i-th data point in the following form:

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1 - h_{ii}}, \quad i = 1, 2, ..., n \tag{9}$$

   where r_i is the i-th studentized residual.

5. (4 points) For the *soft drink delivery time* data set with in total $n = 25$ data points and $p = 2$ features/inputs, we can calculate the leverage score, h_{ii}, as well as the Cook's measure and $DIFFITS$ measure shown in Figure 1. Besides, the average leverage score is $\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = 0.08$. From these results, please identify which points are high leverage points and which points might be highly influential? Please explain your answer. (Simply assume $F_{0.5,2,23} \approx 1$.)

Observation i	(a) h_{ii}	(b) D_i	(c) $DFFITS_i$
1	0.10180	0.10009	–0.5709
2	0.07070	0.00338	0.0986
3	0.09874	0.00001	–0.0052
4	0.08538	0.07766	0.5008
5	0.07501	0.00054	–0.0395
6	0.04287	0.00012	–0.0188
7	0.08180	0.00217	0.0790
8	0.06373	0.00305	0.0938
9	0.49829	3.41835	4.2961
10	0.19630	0.05385	0.3987
11	0.08613	0.01620	0.2180
12	0.11366	0.00160	–0.0677
13	0.06113	0.00229	0.0813
14	0.07824	0.00329	0.0974
15	0.04111	0.00063	0.0426
16	0.16594	0.00329	–0.0972
17	0.05943	0.00040	0.0339
18	0.09626	0.04398	0.3653
19	0.09645	0.01192	0.1862
20	0.10169	0.13246	–0.6718
21	0.16528	0.05086	–0.3885
22	0.39158	0.45106	–1.1950
23	0.04126	0.02990	–0.3075
24	0.12061	0.10232	–0.5711
25	0.06664	0.00011	–0.0176

Figure 1: Statistics for detecting high-leverage and high influential observations for the *soft drink delivery time* data set.

6. (4 points) Robust regression methods can be used to fit the majority of the data points. Please explicitly give out the robust Huber's t function, $\rho(z)$ as well as its first order derivative $\psi(z) = \frac{\partial \rho(z)}{\partial z}$.
(2 points) Briefly sketch the robust Huber's t function $\rho(z)$ as well as its first order derivative $\psi(z)$ in two separate figures. Assume $t = 2$ for simplicity.
(2 points) Explain (briefly) why the Huber's t function is able to achieve robustness against high influential points/outliers as compared to the least-squares criterion function $\rho(z) = \frac{1}{2}z^2$?

7. (6 points) The cost function for solving the robust M-type estimator is given as

$$\hat{\boldsymbol{\beta}}_M = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) \triangleq \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{s}\right), \tag{10}$$

where we assume the scaling parameter s has been pre-selected. We have seen one iterative algorithm called iteratively re-weighted least-squares (IRLS). Alternatively, we could also use the more general gradient descent $\boldsymbol{\beta}^{\eta+1} = \boldsymbol{\beta}^{\eta} + \alpha^{\eta}\boldsymbol{d}^{\eta}$. Suppose we are using Huber's robust criterion function, $\rho(.)$, with $t = 2$, could you please design a proper descent direction, $\boldsymbol{d}^{\eta}$ for the $(\eta+1)$-th iteration, give its explicit mathematical expression, and explain why it is a proper one?

8. (3 points) How to generate a number of n i.i.d. samples, $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$, from a contaminated Gaussian distribution, $p(\varepsilon) = (1-\alpha)\mathcal{N}(\varepsilon; 0, \sigma^2) + \alpha\mathcal{N}(\varepsilon; 0, \sigma_c^2)$, with $\sigma^2 = 1$, $\sigma_c^2 = 10$, and $\alpha = 0.05$. Sketch your idea briefly.

Question 3: Nonlinear Regression (25ps in total)

Bates and Watts [1988] use the Michaelis-Menten model for chemical kinetics to relate the velocity (rate) y of an enzymatic reaction to the substrate concentration x. The model (with noise) is

$$y = \frac{\theta_1 x}{x + \theta_2} + \varepsilon. \tag{11}$$

The data for the velocity (rate) of a reaction in a puromycin experiment are shown in the table and plotted in the figure below.

TABLE 1 Reaction Velocity and Substrate Concentration for Puromycin Experiment

Substrate Concentration (ppm)	Velocity [(counts/min)/min]	
0.02	47	76
0.06	97	107
0.11	123	139
0.22	152	159
0.56	191	201
1.10	200	207

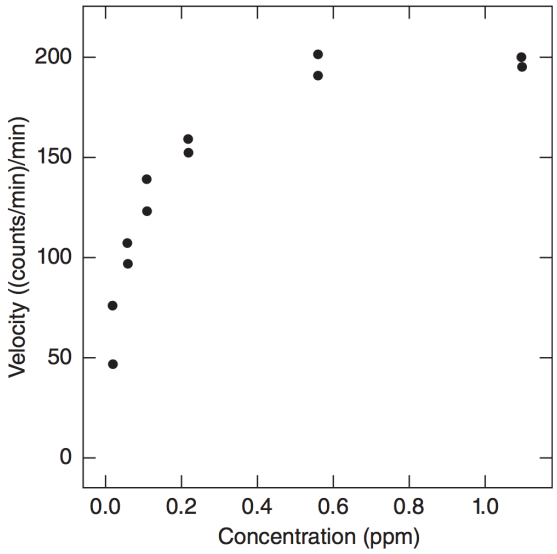

Figure 2: Scatter plot of reaction velocity versus substrate concentration.

Please solve the following problems:

1. (4 points) Michaelis-Menten model is a nonlinear regression model because the mean response function $f(x;\boldsymbol{\theta}) = \frac{\theta_1 x}{x+\theta_2}$, where $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$, is nonlinear in terms of $\boldsymbol{\theta}$. Please calculate the gradient of the mean response function $f(x;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, denoted by $\nabla_{\boldsymbol{\theta}} f(x;\boldsymbol{\theta})$.

2. (2 points) Write out the nonlinear cost function $S(\boldsymbol{\theta})$ in form of "sum of error squared" that you aim to minimize with respect to the model parameters, $\boldsymbol{\theta}$.

3. (5 points) A student F wrote some codes to implement the Gauss-Newton's method according to the following description:

   - Find an initial estimate $\hat{\boldsymbol{\theta}}^0$
   - Select a threshold $\delta_T = 10^{-6}$
   - For $\eta = 0, 1, 2, ...,$ do
     (a) Compute $\hat{\boldsymbol{\theta}}^{\eta+1} = \hat{\boldsymbol{\theta}}^{\eta} + \left(\mathbf{Z}_\eta^T \mathbf{Z}_\eta\right)^{-1} \mathbf{Z}_\eta^T \left(\mathbf{y} - \mathbf{f}_\eta\right)$, where

$$\mathbf{y} - \mathbf{f}_\eta = [y_1 - f(x_1;\hat{\boldsymbol{\theta}}^\eta), y_2 - f(x_2;\hat{\boldsymbol{\theta}}^\eta), ..., y_n - f(x_n;\hat{\boldsymbol{\theta}}^\eta)]^T$$
$$\mathbf{Z}_\eta = [\nabla_{\boldsymbol{\theta}} f(x_1;\hat{\boldsymbol{\theta}}^\eta), \nabla_{\boldsymbol{\theta}} f(x_2;\hat{\boldsymbol{\theta}}^\eta), ..., \nabla_{\boldsymbol{\theta}} f(x_n;\hat{\boldsymbol{\theta}}^\eta)]^T$$

     (b) Compute $\delta_\eta = |S(\hat{\boldsymbol{\theta}}^{\eta+1}) - S(\hat{\boldsymbol{\theta}}^\eta)|$ and compare it with δ_T. Terminate the iterative process if $\delta_\eta < \delta_T$; Otherwise $\eta := \eta + 1$ and proceed with the next iteration.

   (3-a, 2 points) F realized that the program never ends after so many hours (way longer than expected) on a modern computer, what reason(s) would it be? **Open ended.**
   (3-b, 3 points) F also realized that the cost function $S(\hat{\boldsymbol{\theta}})$ is not ensured to decrease with iterations, how to remedy this problem? **Open ended.**

4. (6 points) If the covariance matrix of the random error terms, $\boldsymbol{\Sigma}$ is known and not identical to an identity matrix, the above Gauss-Newton iterations in sub-problem 4 can be modified to be

$$\hat{\boldsymbol{\theta}}^{\eta+1} = \hat{\boldsymbol{\theta}}^{\eta} + \left(\mathbf{Z}_{\eta}^{T}\boldsymbol{\Sigma}^{-1}\mathbf{Z}_{\eta}\right)^{-1}\mathbf{Z}_{\eta}^{T}\boldsymbol{\Sigma}^{-1}\left(\mathbf{y} - \mathbf{f}_{\eta}\right), \tag{12}$$

where

$$\mathbf{y} - \mathbf{f}_{\eta} = [y_1 - f(\mathbf{x}_1; \hat{\boldsymbol{\theta}}^{\eta}), y_2 - f(\mathbf{x}_2; \hat{\boldsymbol{\theta}}^{\eta}), ..., y_n - f(\mathbf{x}_n; \hat{\boldsymbol{\theta}}^{\eta})]^T, \tag{13}$$

$$\mathbf{Z}_{\eta} = [\nabla_{\boldsymbol{\theta}} f(\mathbf{x}_1; \hat{\boldsymbol{\theta}}^{\eta}), \nabla_{\boldsymbol{\theta}} f(\mathbf{x}_2; \hat{\boldsymbol{\theta}}^{\eta}), ..., \nabla_{\boldsymbol{\theta}} f(\mathbf{x}_n; \hat{\boldsymbol{\theta}}^{\eta})]^T. \tag{14}$$

Could you please explain how is the right-hand-side of Eq.(12) derived. Hint: combine the technique applied for deriving the generalized least-squares with the first-order Taylor expansion.

5. (4 points) If the lab researcher has a strong belief that in his data set there is maximally 3 percent of outliers, but the rest of the data are ordinary subject to i.i.d. Gaussian random error $N(0, \sigma^2)$, could you propose a regression equation that possibly achieves higher robustness than the ordinary nonlinear least-squares. Just give the cost function with which we can solve for a robust estimator of $\boldsymbol{\theta}$.

6. (4 points) Can you use another kind of method instead of nonlinear regression to fit the model parameter in Eq.(11)? What are the pros and cons of this method as compared with nonlinear regression.

Question 4: Logistic Regression (15ps in total)

Consider the following binary logistic regression model:

$$y_i = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} + \varepsilon_i. \tag{15}$$

- The output y_i takes on only two possible outcomes, generically "is-the-case ($y_i = 1$)" and "not-the-case ($y_i = 0$)".
- The input $\mathbf{x}_i = [1, x_{i,1}, x_{i,2}, ..., x_{i,k}]^T$ are defined similarly as in the multiple linear regression model.
- The model parameters $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, ..., \beta_k]^T$ are to be estimated.
- The random error is denoted by ε_i.

We assume that the output y_i given a specific $\mathbf{x}_i$ is a Bernoulli random variable with the probability mass function defined as follows:

$$P_i(y_i = 1|\mathbf{x}_i) = p_i = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}, \quad P_i(y_i = 0|\mathbf{x}_i) = 1 - p_i, \tag{16}$$

where p_i is a short-hand notation for the probability.

Please solve the following problems:

1. (2 points) Why don't we use multiple linear regression for the above model? (brief answer 1 reason)
2. (3 points) Please derive the expectation of y_i given $\mathbf{x}_i$, i.e., $E(y_i|\mathbf{x}_i)$.
3. (6 points) Given a dataset $\mathcal{S} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), ..., (y_n, \mathbf{x}_n)\}$, where we assume the n data points are mutually independent. Please show that the log-likelihood function $\ln L(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X})$ is equal to $\sum_{i=1}^n y_i \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^n \ln\big(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})\big)$.
4. (4 points) How would you like to modify the above regression model for a data set in which the output (y) is the grade (A, B, C, D four categories) and the input is a vector of lecture attendance rate (x_1), hours spent on the lecture materials per week (x_2). Please give a modified model, but you don't need to fit the parameters.

Bonus Question: Bayesian Linear Regression (10ps in total)

We consider the Bayesian linear regression model given in the lecture slides, where the prior distribution and the likelihood are given as $p(\mathbf{w}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_p)$ and $p(\mathbf{y}|\mathbf{X}, \mathbf{w}) \sim \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}_n)$, where both $\mathbf{\Sigma}_p$ and σ^2 are known. Using the fundamental results of linear Gaussian system, we can easily derive the joint distribution:

$$p(\mathbf{w}, \mathbf{y}) \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma}_p, & \mathbf{\Sigma}_p\mathbf{X}^T\\ \mathbf{X}\mathbf{\Sigma}_p, & \mathbf{X}\mathbf{\Sigma}_p\mathbf{X}^T + \sigma^2\mathbf{I}_n\end{bmatrix}\right). \tag{17}$$

Please prove (with the aid of the hints below) that the mean of the posterior distribution $p(\mathbf{w}|\mathbf{y}; \mathbf{X})$, denoted by $\bar{\mathbf{w}}$, is equal to:

$$\bar{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X} + \sigma^2\mathbf{\Sigma}_p^{-1}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{18}$$

Hint: If the two random variables $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{y} \in \mathbb{R}^{d_y}$ are jointly Gaussian with the following joint distribution:

$$p(\mathbf{x}, \mathbf{y}) \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix}\right), \tag{19}$$

then it is easy to derive the following conditional probabilities:

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{x|y}, \Sigma_{x|y}) \tag{20}$$

where

$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \Sigma_{xy}\Sigma_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y), \qquad \Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}. \tag{21}$$