Mid-term Exam
March 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: ______________________

Do not use pencil.

Question	SC	Q1	Q2	Q3	Q4	Grade
Mark						

Single Choice Questions: (2 points × 15 questions = 30 ps)

1. Regression analysis is a statistical technique for investigating and modeling the relationship between variables. Is this statement true?
(*a*) Yes, it is true. (*b*) No, it is definitely false.

2. Let random variable $V \sim \chi^2_v$ and random variable $W \sim \chi^2_\eta$. If V and W are also independent, then $\frac{V/v}{W/\eta}$ follows
(*a*) F distribution.
(*b*) non-central $\chi^{2'}_{v,\eta}$ distribution.
(*c*) central $\chi^2_{v-\eta}$ distribution.

3. Scatterplot matrix always reveals the underlying relationship between the output and inputs.
(*a*) True (*b*) False

4. Which of the following statements is correct?
(*a*) In linear regression models, the output y must be linear in terms of both the model parameters and the inputs.
(*b*) A linear parameter estimator $\hat{\boldsymbol{\beta}}$ must be linear in terms of both the input x and output y.
(*c*) The maximum likelihood parameter estimator of the polynomial regression model with one input x and Gaussian i.i.d. random error terms is a linear estimator.
(*d*) None of the above statements is correct.

5. Consider the multiple linear regression model where the random error terms are zero mean, i.e., $E(\boldsymbol{\varepsilon}) = \mathbf{0}$. When using the least-squares parameter estimation strategy, the "hat matrix" is computed by $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, where $\mathbf{X}$ is the design matrix. Which of the following statements concerning the "hat matrix" is correct?
(*a*) $\mathbf{He} = \mathbf{0}$, where $\mathbf{e}$ is the vector of residuals computed as the difference between the outputs $\mathbf{y}$ and the LS fitted values $\hat{\mathbf{y}}$, and $\mathbf{0}$ is a vector of all zeros.
(*b*) $2\mathbf{I} - \mathbf{H}$ is an idempotent matrix; herein $\mathbf{I}$ is the identity matrix having the same size as $\mathbf{H}$.
(*c*) $\mathbf{H}^T = \mathbf{H}^{-1}$.
(*d*) None of the above statements is correct.

6. A polynomial regression model with one input is given by $y = \beta_0 + \sum_{i=1}^{k} \beta_i x^i + \varepsilon$. It is essentially a linear regression model, why?
(a) It is linear in terms of the inputs, x^i, $i = 1, 2, ..., k$
(b) It is linear in terms of the model parameters, β_i, $i = 1, 2, ..., k$
(c) It is linear in terms of both the model parameters and the inputs

7. Using analysis-of-variance (ANOVA) for testing the significance of linear (least-squares) regression, we have the corrected sum of squares SS_T, the regression sum of squares, SS_R and the residual sum of squares, SS_{Res}. Which of the following is correct?
(a) $SS_T = SS_R + SS_{Res}$
(b) $SS_{Res} = SS_R + SS_T$
(c) $SS_R = SS_T + SS_{Res}$

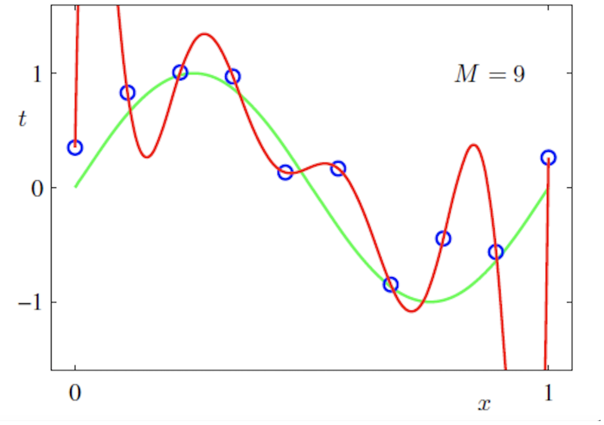
8. When we perform ANOVA for testing the significance of (least-squares) regression, more precisely, testing the hypotheses:

$$H_0 : \beta_1 = ... = \beta_k = 0, \quad H_1 : \beta_j \neq 0 \text{ for at least one } j$$

We need to compute the residual sum of squares, SS_{Res}. What statistical distribution does $\frac{SS_R}{\sigma^2}$ follow under the hypothesis H_1?
(a) central χ^2_k (b) non-central $\chi^2_{k,\lambda}$ (c) central $F_{k,n-p}$ (d) non-central $F_{k,n-p,\lambda}$

9. Suppose we are given a dataset where the output y is stock prices and the input x is time. The data samples were collected on each hour (e.g., 9 AM, 10 AM, 11 AM,...) from 9 AM, 2018.04.27 to 15 PM, 2019.02.26. After fitting a proper regression model, now you are asked to infer the stock prices for (1) 9:30 AM, 2018.12.15; (2) 11:00 AM, 2018.04.18; and (3) 14:00 PM, 2019.04.29. The operations that correspond to the above three scenarios are often called:
(a) interpolation, extrapolation, interpolation.
(b) extrapolation, interpolation, interpolation.
(c) interpolation, extrapolation, extrapolation.
(d) extrapolation, interpolation, extrapolation.

10. The following figure shows the fitting result of a polynomial regression model. The training dataset contains in total 10 data points (see the blue dots). To be precise, the output t is essentially generated from a sinusoidal function (see the green curve) plus some random error, i.e., $t = \sin(2\pi x) + \varepsilon$, where x is the scalar input. The fitted M-th order ($M = 9$) polynomial model is demonstrated by the red curve. What problem is obvious from the figure?



(a) Underfitting
(b) Overfitting
(c) Multicollinearity
(d) No problem at all, a perfect fit!

11. If you agree that there is certain problem in Question 10, please choose a way to alleviate the problem. Otherwise, if you don't think there is any problem in Question 10, please choose (d).
(a) Delete the columns that caused the multicollinearity problem
(b) Increase the order M of the polynomial regression model
(c) Get more data or introduce regularization term in the least-squares estimation
(d) There is no problem in Question 10.

12. Consider multiple linear regression models. Which of the following statements concerning the lack-of-fit test is NOT correct?
(*a*) We have to assume that the random error terms are Gaussian i.i.d. with zero mean and covariance matrix $\sigma^2\mathbf{I}$.
(*b*) To perform the lack-of-fit test, it is better that we have replicate observations of the output y for at least one level of $\mathbf{x}$, but this is not a must, because we could use the data points in a small neighborhood for approximation.
(*c*) SS_{Res} can be decomposed into SS_{PE} and SS_{LOF}, with the former indicating the variance in the data purely due to random error and the latter indicating the variance in the data due to lack-of-fit.
(*d*) A proper scaling of SS_{PE} can be treated as a model dependent estimate on the variance of the random error.

13. When the data shows both curvilinear and periodicity behavior, we should better
(*a*) combine trigonometric and polynomial terms in the regression model with a suitable polynomial order K
(*b*) increase the polynomial model order K
(*c*) use cubic spline model

14. When computing the PRESS residuals, $e_{(i)}, i = 1, 2, ..., n$, we have to fit the model parameters n times, each time based on all data points except for the i-th point, and as a result, the whole computational complexity turns out to be very large.
(*a*) True for linear models
(*b*) False for linear models
(*c*) True for all models
(*d*) False for all models

15. Which of the following statement is correct?
(*a*) When an input takes very large values, it is common to take its logarithm
(*b*) To represent a categorical input variable with m levels, it is common to use one-hot-encoding with m bits
(*c*) The best way of handling missing input value is to complement it with the sample of the corresponding column
(*d*) None of the above statements is correct

Useful Materials

- Theorem: Let $\mathbf{A}$ be a $k \times k$ idempotent matrix of constants with rank p' and $\mathbf{y}$ be a $k \times 1$ multivariate Gaussian random vector with mean $\boldsymbol{\mu}$ and non-singular covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$. Let U be the quadratic form defined by $U = \mathbf{y}^T \mathbf{A} \mathbf{y}$, then $\frac{U}{\sigma^2} \sim \chi^{2'}_{p',\lambda}$, where $\lambda = \frac{\boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu}}{\sigma^2}$.
- Rank of an idempotent matrix is equal to its trace.
- For calculating the first-order derivatives, you may consider to use the formula $\frac{\partial h^T(\mathbf{x}) g(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial h(\mathbf{x})}{\partial \mathbf{x}} g(\mathbf{x}) + \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} h(\mathbf{x})$, where the functions $h(\mathbf{x})$ and $g(\mathbf{x})$ are of appropriate output size.

Question 1: Multiple Linear Regression Model (30ps in total)

A multiple linear regression model is 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 deterministic but unknown model parameters;
- $x_j, j = 1, 2, ..., k$ are the inputs (or features);
- y is the output;
- inputs $x_j, j = 1, 2, ..., k$ are deterministic and precisely known;
- ε stands for the random error.

Given any data set 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;
- $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, ..., \beta_k]^T$ is a $p \times 1$ vector of the unknown model parameters, note that $p = k + 1$ is defined here for simplicity and $p \ll n$;
- $\mathbf{X}$ is an $n \times p$ design matrix of the inputs, which is assumed to be of full rank p;
- $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]^T$ is an $n \times 1$ vector of the error terms.

Please solve the following questions:

1. (6 points) In this sub-problem, let us assume $\boldsymbol{\varepsilon}$ has a known non-constant mean, i.e., $E(\boldsymbol{\varepsilon}) = \boldsymbol{\mu}$ and $Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. For this setup, please derive an unbiased least-squares (LS) parameter estimator of $\boldsymbol{\beta}$ and verify its unbiasedness.

2. (7 points) Following the assumption made on the random error in sub-problem 1, please derive an estimator of σ^2, namely the MS_{Res}. When we further assume that the error terms are Gaussian i.i.d., please show, with the aid of the theorem given in the front, that $\frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-p}$ distribution.

3. (5 points) After the model parameters are fitted, we conduct t-test on each individual parameter estimate, $\hat{\beta}_j$, $j = 1, 2, ..., k$. The results of the hypothesis testing suggest to accept all $H_1 : \beta_j \neq 0$. But interestingly the metric coefficient of determination, R^2 is quite low. What can you conclude from these results?

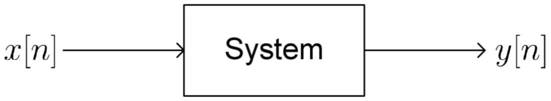
4. (6 points) In this sub-problem, let us assume that the random error terms are independently and identically Gaussian distributed with zero mean and the covariance matrix $Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. For this setup, please derive the covariance matrix of the LS parameter estimator $\hat{\boldsymbol{\beta}}$, i.e., $Cov(\hat{\boldsymbol{\beta}})$. In light of the Gauss-Markov theorem, we know the LS estimator $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE). Please explain the meaning of the terms "linear", "best", "unbiased" respectively.

5. (6 points) In this sub-problem, we assume that the random error terms follow a multivariate Gaussian distribution, more precisely, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Here, we assume $\boldsymbol{\Sigma}$ is a known positive definite covariance matrix but not necessarily to be a diagonal matrix. Please derive the maximum-likelihood estimator of $\boldsymbol{\beta}$, denoted as $\hat{\boldsymbol{\beta}}_{ML}$.

6. **Bonus Question** (5 points) Following the assumption made on the random error in sub-problem 4, we could compute the coefficient of determination, denoted as R^2_A for the current model. Now, let us add one more non-zero input x_{k+1} to the model, re-fit the model parameters, and re-compute the coefficient of determination, denoted as R^2_B for the new model with one more input. Please show that $R^2_A \leq R^2_B$.

Question 2: Linear Least-Squares Applied to System Identification (8ps in total)

System Identification is a methodology for building mathematical models of dynamic systems using measurements of the system's input and output signals. The process of system identification requires that you measure the input and output signals from your system either in the time domain or in the frequency domain. In the following, let us consider a discrete-time linear time-invariant system (see below) with known system input and output, but with **unknown** system impulse response that characterizes the system.



The system model is given by

$$y[n] = \sum_{l=0}^{L-1} h_l x[n-l] + \varepsilon[n], \quad n = 0, 1, 2, ..., N \tag{3}$$

where

- $\boldsymbol{h} \triangleq [h_0, h_1, ..., h_{L-1}]$ are real-valued system impulse response (**unknown**);
- $x[n]$, $n = 0, 1, 2, ..., N$ are real-valued system input at time n;
- $y[n]$, $n = 0, 1, 2, ..., N$ are real-valued system output at time n;
- $\varepsilon[n]$, $n = 0, 1, 2, ..., N$ are real-valued random error terms at time n;
- The first measurement observation time index is 0 and the last observation time index is N, with $N \gg L$.

Note: $x[n]$ with negative integer n is NOT defined. Be careful!

Question: please formulate a multiple linear regression model in matrix form, such as $\boldsymbol{y} = \mathbf{X}\boldsymbol{h} + \boldsymbol{\varepsilon}$ and explain what are your $\boldsymbol{y}$ and $\mathbf{X}$ representing for and give their sizes?

Question 3: Issues with Multiple Linear Models (17ps in total)

Consider the following general multiple linear regression model:

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

where

- $\beta_j, j = 0, 1, 2, ..., k$ are deterministic but unknown model parameters;
- x_k, $k = 1, 2, ..., k$ are standarized inputs (e.g. using unit-length scaling) and they are assumed to be deterministic and precisely known;
- y is the standarized output (e.g. using unit-length scaling);
- ε is the random error.

Given any data set with n data points, the above polynomial regression model can also 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 standarized outputs.
- $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, ..., \beta_k]^T$ is a $p \times 1$ vector of the unknown model parameters, note that $p = k + 1$ is defined here for simplicity and $p \leq n$.
- $\mathbf{X}$ is an $n \times k$ design matrix of the standarized inputs. We let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k]$ where the n-vector $\mathbf{x}_i = [x_{1,i}, x_{2,i}, ..., x_{n,i}]^T$.
- $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]^T$ is an $n \times 1$ vector of the error terms.

Please solve the following questions:

1. (4 points) Explain what is multicollinearity problem and briefly mention the numerical problem it may incur?

2. (5 points) We can use eigensystem analysis to detect multicollinearity problem. Now suppose we are given a data set with $k = 6$ inputs all normalized using unit-length-scaling. The six eigenvalues of $\mathbf{X}^T\mathbf{X}$ are $\lambda_1 = 3$, $\lambda_2 = 1.47$, $\lambda_3 = 0.85$, $\lambda_4 = 0.25$, $\lambda_5 = 0.01$, $\lambda_6 = 0.001$. Please compute the first two largest condition indicies and the condition number as well. What can you conclude from the results?

3. (2 points) Following the sub-problem 2, if we know the associated eigenvector is

$$\boldsymbol{t}_6 = [-0.5, -0.4, -.0.6, -0.8, -0.005, -0.0002],$$

then what linear dependency can you obtain specifically for this example?

4. (6 points) Regularization term $||\boldsymbol{\beta}||_2^2$ can be introduced into the ordinary LS cost function for estimating a better behaved least-squares type parameter estimator, $\hat{\boldsymbol{\beta}}_R$:

$$\hat{\boldsymbol{\beta}}_R = \arg\min_{\boldsymbol{\beta}} S_R(\boldsymbol{\beta}) \triangleq (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_2^2, \qquad (6)$$

where λ is a pre-selected positive number. Please derive $\hat{\boldsymbol{\beta}}_R$ analytically. Please explain why $\hat{\boldsymbol{\beta}}_R$ is helpful for avoiding multicollinearlity problem?

Question 4: Residual Analysis (15ps in total)

Model adequacy checking can be done via residual analysis. In this question, residual analysis is confined to multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{7}$$

where

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

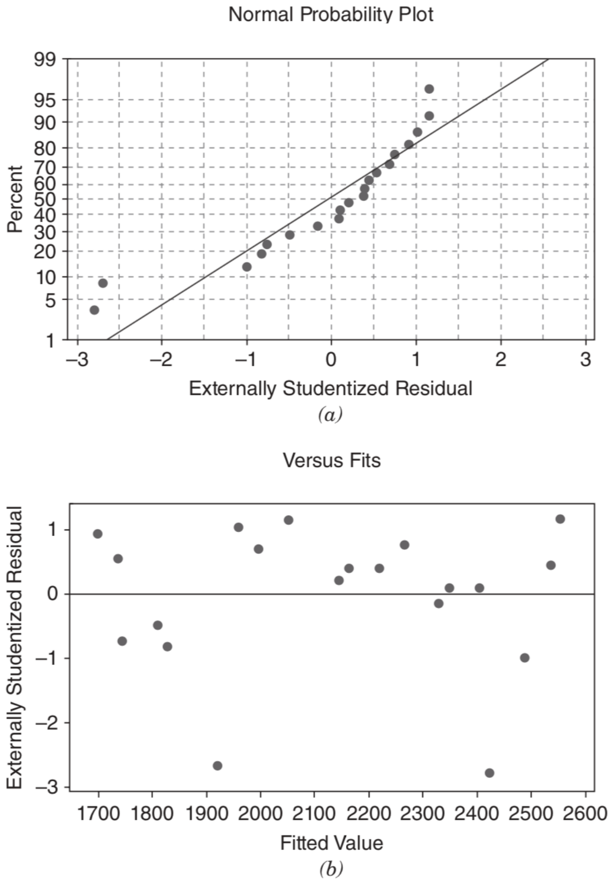
Given any data set with n data points, the above multiple linear regression model can be written in a compact matrix form as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{8}$$

where

- $\mathbf{y} = [y_1, y_2, ..., y_n]^T$ is an $n \times 1$ vector of the outputs.
- $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, ..., \beta_k]^T$ is a $p \times 1$ vector of the unknown model parameters, note that $p = k + 1$ is defined here for simplicity and $p \ll n$.
- $\mathbf{X}$ is an $n \times p$ design matrix of the inputs, which is of full rank p.
- Here, we let $\mathbf{x}_i^T = [1, x_{i,1}, x_{i,2}, ..., x_{i,k}]$ to be the i-th row of $\mathbf{X}$. Thus $\mathbf{X}^T = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$ is of size $p \times n$.
- $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]^T$ is an $n \times 1$ vector of the error terms.

In this question, we aim to test if the assumption “the error terms follow Gaussian i.i.d. with zero mean” holds or not. We narrow down our focus to the least-squares estimator of $\boldsymbol{\beta}$ when computing the residuals.



(a)

(b)

Figure 1: R-student (Externally studentized) residual plots for the rocket propellant data (extracted from one textbook example): subfigure (a) shows the normal probability plot; subfigure (b) shows the residuals versus the fitted values $\hat{y}_i$.

Please answer the following questions:

1. (2 points) Please show that the "hat matrix" $\mathbf{H}$ is idempotent.
2. (9 points) Please write out the explicit form of the R-student (also known as externally studentized) residual. If the random error terms $\boldsymbol{\varepsilon}$ are indeed zero mean Gaussian i.i.d., what statistical distribution does an externally studentized residual follow? Please justify your answer. (Note: just give a brief sketch of the proof.)
3. (4 points) Figure 1 (on top) shows two types of residual plots for the rocket propellant data borrowed from one textbook example. From the figures only, what can you say about the "zero mean Gaussian i.i.d." assumption made on the random error terms $\boldsymbol{\varepsilon}$?