76

ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: Dezhong Xu

Date of Exam: Thursday, February 6, 2025 - Friday, February 7, 2025
Time Limit: 75 minutes
Allowed Materials: None (closed book & closed notes)

Student – NC State University Pack Pledge

I, Dezhong Xu have neither given nor received unauthorized aid on this exam or assignment. I have read the instructions and acknowledge that this is the correct exam.

Student's Printed Name

Dezhong Xu. Feb. 7th 2025

Student Signature *Date*

Exam must be turned in by: _____ DX .

Exam End Time *Student's Initial Agreement*

NOTE: Failure to turn in exam on time may result in penalties at the instructor's discretion.

Exam 1

Please write your answers below each question. You should not have access nor use any materials during this exam.

A reminder that, by taking this exam, you are required to uphold the NC State honor pledge:

"I have neither given nor received unauthorized aid on this test or assignment."

1. In the statistical learning paradigm, we discussed three major goals: statistical inference, predictive modeling, and pattern finding.

   Give a brief real world example for each of these goals. Specify a possible model or method we discussed in class that would help answer the question from each real world example.

- Statistical Inference (4 pts)

— Analyze what possible economic factors impact the country's GDP growth.

show: use parametric approach to build up a function between economic variables (jobless population, first or second industry incomes... such as.) annual history data & historical GDP data. Interpreting the coefficients & p-value for each variable in the model. -1

- Predictive Modeling (4 pts)

— Based on the historical traffic count to predict the future 10 years AADT (annual average daily traffic) for a road segment.

show: use parametric approach to build a SLR model that best fit the historical traffic count which explains the relationship between time & traffic for that segment.

- Pattern Finding (4 pts)

— Distinguish the tumor is benign or not, reveal bad tumor common pattern.

show: Use non-parametric method ('such as KNN method) to classify what common patterns that lead to the bad tumor decision making. (tumor size, paticient age) response ... -2

-3

2. Consider having models characterized by flexibility with the scale going from not very flexible to very flexible.

   a. What type of relationship between flexibilty and squared bias would we expect? Why? (4 pts)

with flexibility increase, the bias would decrease.

The model would be closer fitting the data, and eventually be overfitting

   b. What type of relationship between flexibilty and variance would we expect? Why? (4 pts)

with flexibility increase, the variance would increase.

The model would gradually overfitting the data, so the change of the function w/ different flexibility would be increase.

   c. What type of relationship between flexibilty and training error would we expect? Why? (4 pts)

with flexibility increase, training MSE/error would decrease.

more flexibility would lead to overfitting of the training dataset, so the error will keep decreasing.

   d. What type of relationship between flexibilty and test error would we expect? Why? (4 pts)

with flexibility increase, test MSE/error would decrease then increase.

at first, the variance increase would smaller than the bias decrease so the test MSE would decrease, but after reaching the optimal point (prediction function closest to real function), variance increase > bias decrease, so test MSE increase again.

3. What is a tuning parameter or hyperparameter? How does this differ from a 'regular' parameter in a parametric model? (4 pts)

tuning parameter or hyperparameter is the parameter that would evenly distribute the impact to every coefficient of the variables in the model, hence would direct impact model performance as a whole.

The parameters is the coefficient of each variable in the model, it only explains the contribution relationship between corresponding variable & model output.

-2

-2

5. In the multiple linear regression setting, we discussed a number of model selection methods. State four model selection methods that can be used in the $p > n$ situation. (4 pts)

shrinkage methods

1. LASSO
2. Ridge regression

Dimension reduction. Methods

1. Partial Regression algorithm (PRA).
2. Least Partial Squares.

6. State true or false (no need to explain). (3 pts each)

   a. Ordinary least squares performs variable selection.

   ~~True~~

   b. Ordinary least squares performs shrinkage of coefficient estimates.

   False.

   c. Best subset selection performs variable selection.

   True

   d. Best subset selection performs shrinkage of coefficient estimates.

   False.

   e. Ridge Regression performs variable selection.

   False.

   f. Ridge Regression performs shrinkage of coefficient estimates.

   True

   g. LASSO performs variable selection.

   True

   h. LASSO performs shrinkage of coefficient estimates.

   True.

7. Suppose we have a large data set where we want to perform a regression task. We want to determine the best overall model between a LASSO model and a kNN regression model. We want to use a train test split and compare the best kNN and LASSO model on the test set. We wish to determine the appropriate tuning parameters on the training set only using cross-validation. Fully outline the process for splitting the data, tuning, comparing, and fitting a final overall best model. (10 pts)

1. split the data into training set & test set (70/30)

2. define a tuning grid for both LASSO (B) & KNN (k).

3. for each value in B, implement LASSO method on training set to get the optimal model and run the model on test set to get CV test MSE performance.

4. for each value in K, implement KNN method on training set to get the optimal training model & run the model on test set to get test MSE performance.

5. Pick the optimal LASSO model by selecting the models inside tuning grid B with the lowest test MSE.

   Pick the optimal model KNN model by selecting the models inside tuning grid K with the lowest test MSE.

6. Compare the two models' test MSE to decide the final model

8. Consider the Ridge Regression procedure for fitting a multiple linear regression model. With this model we minimize the following criterion (recall $\lambda \geq 0$):

$$\sum_i \left(Y_i - \beta_0 - X_{i1}\beta_1 - \cdots - X_{ip}\beta_p\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

a. What are the benefits of fitting a Ridge Regression model as compared to an ordinary least squares model? (4 pts)

1. less computational expenserses. It would not need to go through every combination of OLS result.

2. The variance of the caviable would be less, so the optimal model ~~would~~ could have less MSE.

-3

multicol.

b. What happens to our coefficient estimates for a 'large' value of the tuning parameter? What happens for a tuning parameter value near 0? (4 pts)

if λ is large, which shrink the coefficients and decrease the flexibility of the model.

if λ is near 0, the result of the ridge regression would pretty much same as the OLS model.

-3

9. Suppose we fit a multiple linear regression model to data about how much people earn. Our response variable is the `wage` (in 1000's of dollars) and our predictors are `marital_status` (`married`, `never_married`, or `divorced`), and age.

   We fit a linear and quadratic term for age and include an interaction between `marital_status` and age and an interaction between `marital_status` and age squared in the model. Output for the model is given below.

		Estimate	Std. Error	t value	Pr(>|t|)
	(Intercept)	25.293	38.116	0.664	0.507
X_1	marital_statusmarried	-19.780	40.405	-0.490	0.624
X_2	marital_statusnever_married	-31.760	40.992	-0.775	0.439
X_3	age	2.846	1.611	1.767	0.077
X_3^2	I(age^2)	-0.024	0.017	-1.470	0.142
X_1X_3	marital_statusmarried:age	2.024	1.716	1.179	0.238
X_2X_3	marital_statusnever_married:age	2.230	1.820	1.225	0.221
$X_1X_3^2$	marital_statusmarried:I(age^2)	-0.025	0.018	-1.412	0.158
$X_2X_3^2$	marital_statusnever_married:I(age^2	-0.032	0.020	-1.607	0.108

   a. Write down the fitted equation for $\hat{y}$. Define any indicator variables as needed. (4 pts)

$$\hat{y} = 25.29 - 19.78X_1 - 31.76X_2 + 2.846X_3 - 0.024X_3^2 + 2.024X_1X_3 - 2.23X_2X_3 - 0.025X_1X_3^2 - 0.032X_2X_3^2$$

✓

where X_1: marital_status married

X_2 —— unmarried

X_3: age.

   b. One column of the output represents the t-value or t-statistic. What is the usefulness of this t-value? (2 pts)

it is the result of a null hypothesis for each variable, can be turned into p-value to reveal variable's significance.

ok

0

$X_1 = 1 \quad X_2 = 0 \quad X_3 = 30$

c. Write down the form of a predicted value for somone that is `married` and has an age of 30. No need to simplify. (2 pts)

$\hat{y} = 25.29 - 19.78 + 2.846 \times 30 - 0.024 \times 30^2 + 2.024 \times 1 \times 30 - 0.025 \times 1 \times 30^2$ ✓

d. Write down the form of a predicted value for somone that is `divorced` and has an age of 30. No need to simplify. (2 pts)

$X_1 = 0 \quad X_2 = 1 \quad X_3 = 30$

$\hat{y} = 25.29 - 31.76 + 2.846 \times 30 - 0.024 \times 30^2 - 2.23 \times 1 \times 30 - 0.032 \times 1 \times 30^2$

−1

f. Conceptually, what does including an interaction between `marital_status` and age and an interaction between `marital_status` and age squared do to our model as compared to a model without those interactions (that still includes a main effect for `marital_status` and a linear and quadratic term for age)? (3 pts)

The "wage" would fit different slope with "age" variables, and the slope also changes depending on the marital status.

separate quadratics

−2

g. The F-statistic for the global model test is 46.26 on 8 numerator and 2991 denominator degrees of freedom. The p-value for the test is very close to zero.

i. Write down the null and alternative hypotheses for this global test. (3 pts)

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

$H_a: \beta_j \neq 0$ at least one $\beta_j \neq 0$

−1

ii. We see a significant global test but none of the coefficient tests are significant. What do you think could be causing this issue? (3 pts)

The colinearity & correlation among the variables. ✓

h. What type of plot might we look at to investigate the homogenous error variance (i.e. the assumption of equal error variance)? (3 pts)

QQplot or ppplot. →

−7