89

Nice job overall!

DELTA Testing Services

go.ncsu.edu/testing

Campus Box 7555
1730 Varsity Dr.
Venture IV, Suite 236
Raleigh, NC 27695-7113

919.515.1560 phone
919.515.7180 fax

delta-testing@ncsu.edu

NC STATE UNIVERSITY

DELTA Testing Services

Student Name: Matthew Bray **Date:** 2/7/25

Student's NCSU Email Address: rmbray@ncsu.edu

Course: ST 563 601 **Exam #:** 1

Start Time: 1:35pm **End Time:** 2:30pm

Proctor's Name (Print): Alex Khoury

Proctor's Signature:

Institution: Bridgewater State University

PLEASE SIGN & DATE THIS SHEET AND RETURN ALONG WITH THE EXAM

Proctoring Guidelines

If you are unable to comply with the following, please destroy the exam and have the student submit the name of another proctor for approval.

1. Please ask student for their photo ID.
2. **Have the student put their name on the exam and exam answer sheet.**
3. The test should be conducted in an atmosphere conducive to good concentration (quiet, good lighting, etc.).
4. The student must take the exam without outside help. Have the students leave all materials (except blank paper, pen or pencil, or calculator, as needed) outside the testing room. This includes notes, books, calculators, phones, etc. (excluding materials required for the exam).
5. Close and constant supervision must be provided.
6. Please scan and email the proctoring form, completed exam, and any formula sheets permitted for the assessment to delta-testing@ncsu.edu or fax to 919-515-7180.
7. Not including exams that permit all notes or textbooks, students should not be permitted to leave the testing room with formula sheets or scrap paper unless explicitly stated.
8. DO NOT GIVE THE EXAM TO THE STUDENT TO MAIL BACK

If you have any questions, please contact DELTA Testing Services at our main Venture IV location via phone: (919)-515-1560 or e-mail: delta-testing@ncsu.edu.

Thank you for assisting our students.
DELTA Testing Services
NC State University

Updated March 2022

ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: Richard Matthew Bray

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, Matthew Bray 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

Student Signature

07Feb25
Date

Exam must be turned in by: ____________________

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)

Want to understand what variables may impact on outcome, ie does car color influence sales volume, as compared to other variables.

linear model ok

- Predictive Modeling (4 pts)

Predict patient response to a therapeutic based on known biological variables for the patients. k-nearest ~~neighbor~~ neighbor, or supervised models in general

- Pattern Finding (4 pts)

clustering varibles, perhaps transactions in a worldwide financial system. ok

Unsupervised learning methods.

model: clusterng

0

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)

   Decrease in flexibility results in higher squared bias. The model moves further from the observations.

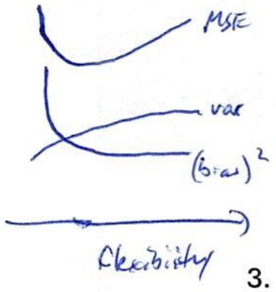
   b. What type of relationship between flexibilty and variance would we expect? Why? (4 pts)

   Variance increases as flexibility increases. The model has to "move" more in order to flex closer to the observations.

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

   Training error decreases as model flexibility increases. The model can go closer to, and learn, about small variations in the data as it becomes more flexible.

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



   ~~Test error both~~ ~~[illegible]~~ Test error will generally be higher than training error, and in general will have a "U" shape relative do flexibility. This is because the bias has an initial drop, but variance will increase.

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

A hyper-parameter allows for changing the "base" of the model. For example in K-nearest neighbors, it allows for averaging across different K-numbers of neighbors. In Ridge, LASSO, and Elastic Net models, they allow for shrinkage of the model parameters to get smaller models with more model parity. Hyperparameters are not statistical parameters upon which inference can be performed. K isn't either!

−1

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)

Elastic Net ✓

LASSO ✓

[illegible]

Best ~~Subset Selection~~

Forward Stepwise Selection ✓

-1

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

   a. Ordinary least squares performs variable selection.

      False ✓

   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 ✓

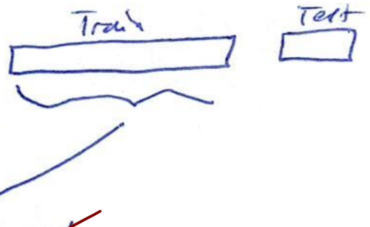
   h. LASSO performs shrinkage of coefficient estimates.

      True ✓

-1

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 Data into training/test split (80/20, 70/30) using SRS without replacement. ✓



2) create k-folds of the data using SRS without replacement.

3) **KNN**

create grid of tuning parameters (k). Fit model at each value of k ~~on each~~ fold, then combine metrics for each value of k. Select value of k from model with best metric. ✓

LASSO

create grid of tuning parameters (λ). Fit model at each value of λ on each fold, then combine metrics for each value of λ. select value of λ with best metric, or best λ + 1se.

-2

4) For each model, predict on test set. -1

fit to what data?

5) Fit model with best prediction metric on full dataset. ✓

-3

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)

The predictors shrink, reducing the model variance. Also better at prediction.

also multicollinearity issue

−1

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)

Large tuning, estimate shrink ✓

Small tuning, estimates are same as OLS.

−1

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
marital_statusmarried	-19.780	40.405	-0.490	0.624
marital_statusnever_married	-31.760	40.992	-0.775	0.439
age	2.846	1.611	1.767	0.077
I(age^2)	-0.024	0.017	-1.470	0.142
marital_statusmarried:age	2.024	1.716	1.179	0.238
marital_statusnever_married:age	2.230	1.820	1.225	0.221
marital_statusmarried:I(age^2)	-0.025	0.018	-1.412	0.158
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)

-1

$$\hat{y} = 25.293 - 19.780\,x_{married} - 31.760\,x_{never\text{-}married} + 2.846\,x_{age} - 0.024\,x_{age}^2$$
$$+ 2.024\,x_{married}x_{age} + 2.230\,x_{never\text{-}married}x_{age} - 0.025\,x_{married}x_{age}^2$$
$$- 0.032\,x_{never\text{-}married}x_{age}^2$$

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

Create confidence intervals around the estimates

-1

-2

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)

$$\hat{y} = 25.293 - 19.780 + 2.846(30) - 0.024(30)^2 + 2.024(30) - 0.025(30)^2$$

✓

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)

$$\hat{y} = 25.293 + 2.846(30) - 0.024(30)^2$$

✓

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)

allows ~~use~~ us to see how the effect of marital_status changes as age changes and how the effect of age changes as marital_status changes.

true generally for interactions being included

Here we actually fit separate quadratics for each marital status

−1

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)

$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$

$H_A: \beta_1 = \cancel{\beta_2 = \dots} = \beta_p \neq 0$

at least one β_j

−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)

✓

Overfitting of the model. The model may ~~fit~~ be flexible and fit the overall data well, but to many predictors involved for any one to be important alone.

tests are for $\beta_j = 0$ after accounting for other x_j's

−1

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)

residuals plotted by fitted results ✓

−3