76

ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: Alexander Devoid

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, Alexander Devoid 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 2/6/25 *Date*

Exam must be turned in by: 11:30 *Exam End Time* *Student's Initial Agreement*

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

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: Alexander Devoid **Date:** 2/6/2025

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

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

Start Time: 10:15 am **End Time:** 11:30

Proctor's Name (Print): Jessica Snow

Proctor's Signature: Jessica Snow

Institution: Southwestern Community College

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

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)

Simple Linear Regression - We ~~want~~ to understand better if students who spend more time studying get better grades. SLR is a Parametric model, easy to interpret and explain the relationship between hours studied and test scores. We can use SLR to estimate how much how each hour spent studying impacts test score.

- Predictive Modeling (4 pts)

non-Parametric - Deep learning - we want to predict the winner of a digital horse racing game. We have data on all the horse races, along with metadata on all horses. We have access to lots of compute and only care if the predictions are better than a coin toss. We dont need to explain why certain horses win.

- Pattern Finding (4 pts)

~~K~~nn - we have data on newspaper readers. we know lots about each reader and we want to categorize these readers into profile groups that may help us use our marketing budget.

-1

-1

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)

   A flexible model will have low bias.

   A not-very flexible model will have high bias due to the assumptions of the model.

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

   A mode with more flexibility would fit traing dat more exact on data with high variance.

   ? -2

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

   A model with more flexibility will have a lower training error due to the more exact fit,

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

   A very flexible model risks being overfit, with a High test error. -2 u shape

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

   A tuning parameter (lik k in KNN) is set by the analyst as an arbitrary value that yields best results. A regular Parameter is found in the data and the analyst has no control over what it is, as it was observed, not set or created.

-4

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)

Forward StcPwise ✓ −2
Backward StcPwise
Best Subset
LassO (reduces some B_s to zero) ✓

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 (may shrink coefficient variable to zero) ✓

   h. LASSO performs shrinkage of coefficient estimates.

   True ✓

−2

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)

- Create a matrix of possible tuning parameter values
- train/test split
- using CV split data into folds so that each fold gets a chance to be the test data.
- for each value in the tuning parameter matrix run the K fold cross validation process.
- asses each average test error for each CV process, or another model performance metric i.e RMSE, use the tuning param that performed the best.
- do this for the Lasso and Knn models.

Compare model metrics between the models with their optimized tuning parameter.

how? fit on full training data

test on test set, pick best, fit final model on full data set

-1 -3 -5

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

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

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

Ride Regression Penilizes Coeficents that dominate.

-2

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)

a tuning Parameter of zero would just be the OLS.

A landa that's high risks over modifying Coef. estimates, ok

-2

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)

$$\hat{y} = 25.293 + -19.78(X_1) + -31.76(X_2) + 2.846(X_3) + -0.024(X_3^2) + 2.024(X_1 \cdot X_3) + 2.23(X_2 \cdot X_3) + -0.025(X_1 \cdot X_3^2) + -0.032(X_2 \cdot X_3^2)$$

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

The t distribution shows all possible t values if the null hypotheses is true and the relationship is due to random chance. We generally want to see a t value 2 standard deviations away from the mean of the t distribution to reject the null hypotheses.

ok

0

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)

$Y = 25.93 + -19.78(1) + -31.76(0) + 2.846(30) + -.024(30^2) + 2.024(1 \cdot 30) + 2.23(0 \cdot 30) + -0.025(1 \cdot 30^2) + -0.032(0 \cdot 30^2)$

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)

Same as above but all X_1 and X_2 values are zero.

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)

It gives unique slopes to age for each marital status group.

separate quadratics −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)

−3 The null hypotheses is that the interaction term have no effect on relationship

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 interaction terms improve the model but the model at baseline does not reject the null hypotheses. −3

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)

−3 Q Q Plot - If points on the plot are in a straight line then the error variance is normal. If the points are not in a straight line then assumptions about the data may be incorrect or transformations should be applied. −10

The t distribution shows all possible t values if we assume the null hypothesis is true and the relationship is due to random chance. We generally want to see a t value 2 standard deviations from the mean to reject the null hypothesis and