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

91
well done!

**Student Name:** David Grant **Date:** 2/7/2025

**Student's NCSU Email Address:** dgrant@ncsu.edu

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

**Start Time:** 9:36 **End Time:** 10:51

**Proctor's Name (Print):** Dylan F. Switala

**Proctor's Signature:** Switala

**Institution:** Ramapo 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

# ST 563 601 – SPRING 2025 – POST
# Exam #1

**Student's Name:** David Grant

**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, David Grant 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*

David Grant 2/7/25

*Student Signature* *Date*

**Exam must be turned in by:** 10:47 DG

*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)

Using sample data to make inferences about a population, i.e. taking a survey of 50 people's favorite ice cream flavor and making an inference on the proportion of all people who's favorite flavor is chocolate. A method that can be used to calculate this is a confidence interval. not a model −2

- Predictive Modeling (4 pts)

Given sample data, building a model such as a simple linear regression model to predict the response value given the value of the predictor. i.e. using a baseball player's home runs to predict their batting average. ok

- Pattern Finding (4 pts)

Again using sample data, trying to explore relationships of sample data, especially between predictor variables in a MLR model. A method for this could be comparing different models that include an interaction term or not. i.e. using a player's home runs and hits to predict batting avg, there may be patterns between HR and hits.

−4

−6

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)

   We'd expect an inverse relationship meaning as a model gets more flexible its bias decreases. This happens because as we get more flexible, we are fitting to the sample data better, covering more of the patterns in the data.

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

   We'd expect a direct relationship meaning as flexibility increases so does variance. This is because if were fitting to the sample data more closely, then each observation has more weight in fitting the model, inherently making it more variant.

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

   Inverse. As flex increases, training error decreases. If we're fitting the points more closely, then each observation should be closer to the model prediction.

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

   Test error would decrease initially, but eventually increase as we become very flexible. This is because we may start to overfit to the training data, and not account for variability in new data.

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

   It's a parameter that changes the fit of the model, and does not follow a specific equation. Example k in KNN. Ideally, we'd want to optimize this parameter to produce the lowest bias-variance combination. A regular parameter is directly calculated from the equation and doesn't need to be manually optimized.

10

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)

Best subset selection
forward selection
backward selection
LASSO

-2

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

-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)

Split data into 70/30 or 80/20 train/test sets ideally.

Use 5 or 10 fold CV on training set by splitting training into 5 or 10 equally sized groups. Train for a specific value of the hyperparameter on all but 1 of the groups, and test on the remaining group. Repeat this process so that every group gets to be the test group and calculate the total test error for all groups.

Repeat this for all values of the hyperparameter and see which value produces the lowest test error.

Once we get optimal tuning params (1 for the LASSO, 1 for the kNN), fit each on the entire training set, and test on the test set. (using the optimal tuning vals!)

Whichever model has the lowest test error from the test set is our overall best model. We can then fit this model on the entire initial set to get ready to predict on new data.

9

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)

Ridge regression may be able to shrink coefficients of predictors that are not important so we don't have to worry about overfitting to the model. It can also address any issues with multicollinearity between predictors.

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)

For a large value of the tuning parameter, we will have more "penalty", meaning the coefficients will get closer to 0 and not be as important in the model. Less variance but more bias.

conversely, a tuning param near 0 means coefficients won't shrink as much, so they will still be important, and be similar to OLS model. More variance but less bias.

0

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)

$X_1$ = married = 1, other = 0 $\quad X_2$ = not married = 1, other = 0 $\quad X_3$ = age $\quad X_4$ = age^2

$$\hat{y} = 25.293 + (-19.780)X_1 + (-31.760)X_2 + 2.846X_3 + (-0.024)X_4 + 2.024X_1 \cdot X_3 + 2.230X_2 \cdot X_3 + (-0.025)X_1 \cdot X_4 + (-0.032)X_2 \cdot X_4$$

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

We can calculate the probability of being greater than the t value, which allows us to determine if the variable is significant or not.

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)

$\hat{y} = 25.293 - 19.780 + 2.846(30) - 0.024(30^2) + 2.024(1 \cdot 30) - 0.025(1 \cdot 30^2)$ ✓

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)

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

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)

Including the interaction between marital status and age, as well as mar. status and age$^2$ allows us to see any masking effects that could've been happening w/o them. Meaning age could impact marital status.

−1 ok, but separate quadratics here

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)

$h_0: b_1 = b_2 = b_3 = b_4 = b_5 = b_6 = b_7 = b_8 = 0$

$h_A$: at least one term $\neq 0$

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 terms could be too correllated with each other, especially since we have age$^2$. This can make it seem like none of them are important. ok

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)

A residual plot and look for a trumpet shape. ok

1 vs what?

−1