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: Matthieu Cartron **Date:** 2/7/25

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

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

Start Time: 1002am **End Time:** 1117am

Proctor's Name (Print): Bilha Lucero

Proctor's Signature: Bilha Lucero

Institution: Univ. of New Mexio Testing + Training Ctr.

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

Matthieu
Cartron

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)

Multiple linear regression model in which we aim to model income as a function of education, age, and IQ. Here we are looking to examine if any of these covariates are meaningful in our model. We want to answer questions like:

Is age related to income?

- Predictive Modeling (4 pts) (soccer reference)

KNN model in which we aim to model the # of goals scored by a team in a given game as a function of passes completed, expected goals (xg), and possession. We are looking to try and make accurate predictions regarding the number of goals scored (as accurate as possible) in games that have not yet been played! We are not so interested in which covariates are statistically significant.

- Pattern Finding (4 pts)

Principle components analysis to try and group U.S states that have similar subscriber attrition at two time intervals (unsupervised learning method). We want to see if our PCA finds patterns via grouping - perhaps some states share similar attrition rates across these metrics.

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)

   (inverse relationship)

   As flexibility increases, we would expect the squared bias to decrease as the model fits the data more and more closely.

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

   As flexibility increases, we would expect variance to increase (and vice-versa) more flexibility means that changing a single observation can have a larger impact on our model fit, and this is what our model variance measures.

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

   As flexibility increases, training error decreases because our model more closely fits the training data (trading off a decrease in squared bias for a smaller increase in variance).

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

   As flexibility increases on our trained model, we would expect our test error to increase. As a model becomes more flexible, it begins to overfit to the training data and thus does not generalize as well to the test set.

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

A regular parameter in a parametric model is derived via a closed-form solution (think OLS), whereas a tuning parameter is not arrived at via closed-form solution. It must be "tuned" (a tuning parameter will have some kind of 'search grid), that is, we must test the performance of our model with a variety of candidate values for our tuning parameter to find the optimal value (think k in a KNN model)!

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)

   - Ridge Regression
   - LASSO
   - other higher dimensional regression models
   - Partial least squares — PCA regression

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

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. Train/Test Split (split data into training and test sets, w/ about 70%-80% of data in the training set).
2. Split Training set into v-folds
3. Create a search grid for each model type
    1. For KNN, choose some vector of Candidate values for k
    2. For LASSO, choose some vector of Candidate values for λ.

record CV error

4. For each value of the given hyperparameter, fit the model on v-1 folds, and evaluate the performance on the fold that was left out. Do this repeatedly until all v folds have been left out once (the left out fold is the internal training validation set for that iteration). Average all v CV errors for that value of k (or λ if we are tuning our LASSO Model).
5. Once the above step has been done for each Candidate value of the hyperparameter, choose the hyperparameter that achieved the lowest CV error.
6. Refit both models on the entire training data, each with their respective best choice for k and λ.
7. Evaluate the performance of these tuned models against the test set. Choose the model with the lowest test set error as the final overall best 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)

- Ridge Regression primarily helps us deal w/ multicollinearity, shrinking correlated variables towards one another (so that their effect is distributed between them and is not siloed into only one of them).

- Ridge Regression allows us to tune our model via λ so that we may have a better chance of finding the right balance between $Bias^2$ and Variance (and thus make better predictions).

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)

$$\lambda \sum_{j=1}^{p} \beta_j^2 \leq t$$

As $\lambda \uparrow$, our β values shrink depending on their level of importance to the model (though they will not shrink to zero). Correlated variables will have βs that shrink toward one another.

As $\lambda \downarrow$, our β values grow, and when $\lambda \approx 0$, we effectively have the OLS solution because our penalty is effectively gone.

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)

$$\hat{Y}_{wage} = 25.293 - 19.780X_1 - 31.760X_2 + 2.846X_3 - 0.024X_3^2$$
$$+ 2.024X_1X_3 + 2.230X_2X_3 - 0.025X_1X_3^2 - 0.032X_2X_3^2 + \varepsilon$$

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

This t-test test the hypothesis that the partial effect of a given partial slope (in the MLR case) = 0. The t-value tells us whether we fail to reject this null hypothesis or whether we should reject it (in favor of $H_A: \beta_j = 0$ for a given j)

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)

$$\hat{y} = 25.293 - (19.780 \cdot 1) - \underbrace{(31.760 \cdot 0)}_{\text{cancels}} + (2.846 \cdot 30) - (0.024 \cdot 30^2)$$
$$+ (2.024 \cdot 1 \cdot 30) + \underbrace{(2.230 \cdot 0 \cdot 30)}_{\text{cancels}} - (0.025 \cdot 1 \cdot 30^2) - \underbrace{(0.032 \cdot 0 \cdot 30^2)}_{\text{cancels}}$$

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)

$$\hat{y} = 25.293 - \underset{\text{Cancels}}{\cancel{(19.780 \cdot 0)}} - \underset{\text{Cancels}}{\cancel{(31.760 \cdot 0)}} + (2.846 \cdot 30) - (0.024 \cdot 30^2)$$
$$+ \underset{\text{Cancels}}{\cancel{(2.024 \cdot 0 \cdot 30)}} + \underset{\text{Cancels}}{\cancel{(2.230 \cdot 0 \cdot 30)}} - \underset{\text{Cancels}}{\cancel{(0.025 \cdot 0 \cdot 30^2)}} - \underset{\text{Cancels}}{\cancel{(0.032 \cdot 0 \cdot 30^2)}}$$

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)

It allows our model to potentially capture non-additive patterns covariates (non-linear in X's not βs !). For example, the effect of age on how much someone earns may depend on that person's marital status. A purely additive model would not account for this.

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)

$H_0: \beta_1 = \beta_2 = \ldots = \beta_9 = 0$

$H_A:$ At leest one of $\beta_1, \beta_2, \ldots, \beta_9 \neq 0$

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)

A number of things could be causing this:

- collinearity
- type II error among t-tests
- Overall significance may be achieved but our covariates may not be properly specified - we may want to include fewer terms and only include our main effects

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)

We would want to look at a scatterplot of (standardized) residuals vs their fitted values to assess homogeneity of variance.