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

**Student's Name:** LiLian Ngonadi 200582411

**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, LiLian 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* *DATE* 7/02/2025

## 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)

This is a classical problem, where we get hypothesis testing, confidence interval & prediction interval

Check the last page

- Predictive Modeling (4 pts)

Here we have a response and we are interested in predicting the output given one or more input variable. This can be done using regression model.
Eg To predict the performance of students in an exam

- Pattern Finding (4 pts)

We are interested in understanding the pattern of the data, this can be done using classification, dimension reduction
We used K-nearest model in class to group : Check last page for continuation

less flexible - flexible

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 would expect less variance but an increase in squared bias

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

    we would expect less variance if moving from less flexible to flexible

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

    we would expect the training error to increase and this is because the model becomes too complex and begins to overfit the model.

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

    The test error will decrease but begin to increase again as the model can no longer capture the model complexity leading to overfitting

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

Tuning parameter is used in non parametric model to help reduce variability and help in selecting the best model to avoid overfitting the 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)

① Lasso
② Ridge
③ Backward selection
4) forward selection
5) Best subset selection

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)

· Split the entire data set into training and testing set (the training set can be 70 & test 30 or 80% training & 20% testing)
Then we further split the training set into internal training and validation set
Then for each of the model using the internal training set
i) for lasso using the penalty parameter and this is we carry out cross validation and this is done for each fold where each fold (for example if we are fitting CV=5 we will use four fold as the training set and the remaining as the test set, we will repeat this using different fold as the test set and the remaining as the training set
Then evaluate the model using the test set for each fold and then compare the metrics

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 benefit of fitting a ridge regression is that we will have less variance in the model as compared to the OLS since it shrink its estimate to 0

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 it will shrinkage estimate close to zero

The tuning parameter value near 0 can approximate to the OLS model and also

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)

for married

$wage = 25.293 - 19.780 + 2.846 age$
$- 0.024 age^2 + 2.024 age +$

$x_1 = \begin{cases} 1 & \text{if married} \\ 0 & \text{if otherwise} \end{cases}$

$x_2 = \begin{cases} 1 & \text{if never married} \\ 0 & \text{if otherwise} \end{cases}$

Check the last page

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 value can be used to compare the value to the tabulated value where the $|t_{value}| > t_{tab}$ means we reject the null hypothesis. And is used to compare the model performance of each estimate

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)

$$wage = 25.293 - 19.780 + 2.846(30) - 0.024(30^2) + 2.024(30) - 0.025(30^2)$$

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)

$$wage = 25.293 + 2.846(30) - 0.024(30^2)$$

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)

It changes the interpretability of the estimate. Also it makes the model more flexible and it now has more predictors in the model.

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)

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

$H_1$: at least one of the $\beta_j$'s is not equal to 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)

This could be as a result of multicollinearity in 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)

we could look at the residual plot

9a)

$x_1 = 1$ if married
$\quad\;\; 0$ otherwise

$x_2 = 1$ if never married
$\quad\;\; 0$ otherwise

for married

$$wage = 25.293 - 19.780 + 2.846\,age - 0.024\,age^2 + 2.024\,age - 0.025\,age^2$$

for never married

$$wage = 25.293 - 31.760 + 2.846\,age - 0.024\,age^2 + 2.230\,age - 0.032\,age^2$$

for divorced

$$wage = 25.293 + 2.846\,age - 0.024\,age^2$$

1a) Inference - If i am interesting in checking the average mean age for all the student in ST563, I can use confidence interval, hypothesis testing

ii) Prediction -
If we want to check the performance of students using their grade as the response variable and their reading time, sleeping hours as the predictor variable
- This can be done using regression analysis

iii) Pattern finding
If we want to understand regions that are colder or hotter, we can group them into k groups using KNN

Continuation of 7a

After getting different MSE values, we then select the optimal hyperparameter

we also do the same method for KNN using a grid of K

Then the final model for the KNN and the Lasso can be selected by fitting the entire data set (using the ~~e~~ entire training and testing set).

Then fit the final model on the entire dataset.