ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: Sanith Rao

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, Sanith Rao 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 02/07/2025 *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)

This case is used for example when we want to see whether salary and education level have a correlation. Can be done w. linear regression.

- Predictive Modeling (4 pts)

Example: when you have a bunch of variables affecting selection in a team. Then you have a new player and based on what their variables are, predict whether they make it. Linear regression again.

- Pattern Finding (4 pts)

unsupervised learning methods work well here because they are able to capture complex relationships. Eg. finding out if there are patterns in number of Covid cases in an area

and if there's any underlying patterns there.

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 would have low square bias since it is able to accomodate nearly all data points resulting in low bias.

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

    A flexible model will have high variance since its fit is determined by the data available to us and so it can give different models w. high variance

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

    They are inversely correlated since a flexible model can accomodate all data points and thus will have low test error

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

    We cannot determine any relationship between the flexibity and test error since that value would change depending on the data set we're using.

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

A hyperparameter is not a parameter that directly affects the response variable. Rather it is a new parameter added to enhance the accuracy of the model. This is common in

shrinkage methods like Ridge and Lasso Regression.

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)

The 4 model selection methods are best subset, forward selection, backward selection, and Lasso 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)

First, we want to split the data into a test an training set (80-20 or 70-30). Then we want use just the training set and split it into n folds. Leave one fold and run LASSO for different values of λ. Find the MSE or RMSE as this the metric we use to determine what value of λ we want. Once we find the best λ, this is the λ we use. For KNN, we do the same where we split training set into n folds and leave one out and perform KNN for different values of K. Once we find the K and λ, run both models on the entire training set to find the model with the best fit using RMSE or MSE. Once we determine the best model, then we fit the entire data on it.

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 benefits of a Ridge regression are:
- It useful for shrinkage and finding optimal B coefficients.
- It ensures that outliers get punished.
- Since we scale all the predictors, it ensures that large values of predictors does not sway the data.

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 close to 0 implies that the ordinary squares and ridge model are essentially the same. A 'large' value for a tuning parameter indicates that it has performed a large amount of shrinkage and is now better than just the OLS model.

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)

$\hat{y} = 25.293 - 19.78x_1 - 31.76x_2 + 2.846(age) - 0.024(I) + 2.024(age)x_1 + 2.23(age)x_2 - 0.025(I)x_1 - 0.032(I)x_2$.

x_1 is an indicator variable for married & for not married x_1 is 0, x_2 is 0, then person is divorced

   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 is significant because to tells us whether to accept the null hypothesis or not (from Pr(>|t|)).

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)

$\hat{y} = 25.293 - 19.78 + 2.846(30) - 0.024(900) + 2.024(30) - 0.025(900)$

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)

$\hat{y} = 25.293 + 2.846(30) - 0.024(900)$

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)

I think including these interactions is very helpful in reducing collinearity because there is definitly a lot of correlation b/n age and marital status.

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)

Null hypothesis: The age and marital status have strong correlation.

Alternate hypothesis: Age and marital status have no correlation

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 because global tests look at the data as a whole and is giving a significant value based on that but the coefficient tests capture true, more intricate relationships.

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)

QQ plot