ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: Naman Pujani

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, Naman Pujani 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 2/7/2025
DATE

Exam must be turned in by: 3:47 NP
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 would be when we have some response variable we're trying to measure based on some predictors. Supervised learning & regression tasks are things we've discussed to answer questions from these examples. We also do hypothesis testing on our predictors for our response.

- Predictive Modeling (4 pts)

When we have a regression model we've made and we want to use it to predict future observations. We've used bootstrapping or CV to fit our predictive models and find future observations.

- Pattern Finding (4 pts)

This is when we are trying to find patterns in the data rather than have a response we're trying to measure. Unsupervised learning or non-parametric models would help us answer the question

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) If we have higher flexibility, we'd have lower squared bias. Flexible models tend to capture complex relationships well so it'll be easier for our models to be closer to the truth, having a low squared bias.

   b. What type of relationship between flexibilty and variance would we expect? Why? (4 pts) We'd have higher variance if we have more flexibility. When we overfit our models, they don't generalize well to newer data and this would increase variance

   c. What type of relationship between flexibilty and training error would we expect? Why? (4 pts) Having flexibility would improve our training error. When we train our data too closely to the data its fit on, it improves our training error because it follows patterns too closely on the train set.

   d. What type of relationship between flexibilty and test error would we expect? Why? (4 pts) Having flexibility won't improve our test error. Since our model won't generalize well to new data, it won't capture the test data relationships well, increasing the test error.

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 is sort of an adjustable setting we use to choose our model (ex: model size). We can use these parameters in model selection methods such as train/test/split, CV, holdout method, etc.

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)

- LASSO
- OLS
- Ridge Regression
- Elastic Net

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.

   False

   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)

We split our data into 5 folds and we'll use each fold as the train set. From our training sets, we will pick our hyperparameters. For LASSO, we'll create a grid of our penalty (λ). We will train our model on each fold one by one for our tuning parameter. After taking our sample, we will use our out of bag obs. to predict on unseen data. We will fit our model based on the tuning parameter that gives us the lowest test error. We will repeat this process B times and obtain the different test errors for each tuning parameter on the folds and find the lowest test error.

For knn, we'll create a grid of k values. We will train our data on each fold for each k we are interested in. We will get the lowest test error & fit our model. We'll repeat this process for every k till we get the lowest test error.

We will look at the test errors for each model and find an overall lowest test error for the two models. The model w/ the lowest test error will be fit to our data and that will be our 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 helps w/the shrinking of coefficients. This allows us to have lower variance since some of our coefficients will go to 0 as compared to OLS fit.

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, our coefficients will shrink and some will go to 0. For a tuning parameter value near 0, we will have something close to the OLS Fit.

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)

$$\hat{y} = 25.293 - 19.78X_1 - 31.76X_2 + 2.846(age) - .024(age^2) + 2.024(age)X_1 + 2.230(age)X_2 - .025(age^2)X_1 - .032(age^2)X_2$$

$X_1 = \begin{cases} 1, & \text{married} \\ 0, & \text{o.w} \end{cases}$, $X_2 = \begin{cases} 1, & \text{never married} \\ 0, & \text{o.w} \end{cases}$

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

It allows us to calculate p-values that tell us whether or not an estimate is significant based on the model that were fit

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)

$\hat{Y} = 25.293 - 19.78 + 2.846(30) - .024(30^2) + 2.024(30) - .025(30^2)$

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)

$\hat{Y} = 25.293 + 2.846(30) - .024(30^2)$

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 provides us with different slopes for age. This allows us to understand the effect of age on the diff levels of marital status and whether or not it's significant 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)

$H_0: \beta_0 = \beta_1 = ... = \beta_j = 0$ vs

H_A: at least one β differs

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)

It's possible that there is a different interaction in the model that we didn't fit which could be significant or the sums of squares is different for a model w/interaction compared to the sums of squares for a model w/out interaction

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)

Standardized Residuals Plot