ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: Zach Binder

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, Zach Binder 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* 02/07/2025

Exam must be turned in by: 11:24 *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)

Ex: Determining if we can conclude whether there is statistically significant evidence the pop prop for influenza vaccine usage is greater than 50% given a sample

model: hypothesis test of proportion

- Predictive Modeling (4 pts)

Ex: predicting a person's income using graduation status, age, and marital status

model: mutiple linear regression

- Pattern Finding (4 pts)

Ex: Might there be different species of trees in a 10mile by 10mile grid based on clustering

model: Spatial point pattern/process

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)

    As flexibility increases, squared bias decreases as the model more closely follows the noise of the data and more accurately replicates the pattern of that dataset (though not necessarily the pattern of the pop)

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

    As flexibility increases, variance will increase as the model is more closely following the underlying noise of the model. This will cause the variance to be high.

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

    As flexibility increases we would expect training error to decrease as the model would more closely follow the training data so values like residuals would be small.

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

    As flexibility increases initially test error will decrease as the model will more accurately model the data, but eventually the test error will increase as the model flexibility will cause overfitting to the training dataset

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 something we choose which impacts the flexibility/fit of the model. It is different than a regular parameter in a parametric model (like μ) as that is a population metric that is fixed and unknown that we are trying to predict normally by fitting a model and using a tuning parameter

We use tuning parameters to come up with what we believe to be the best model based on our training dataset

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
   - Ridge Regression
   - Elastic Net
   - Best subset selection
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.

      True

   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 will split our data into a training and test sets (normally 80% to 20%) (train, test)
- For LASSO we will create a tuning grid of parameters (λ, α) TO use cross validation, we will split the training dataset into v folds. We will train our LASSO model using the tuning grid by using one possible (α, λ). The training will happen on v-1 folds, with the test occuring on the fold left out of the v-1. we will get a test error. We will repeat for that (α, λ), such that every fold is used once as a test fold. We will combine test errors to get a total test error average.

  We repeat this for all (α, λ)'s in the tuning grid. We choose the (α, λ) with the lowest test error as our best model.

- we repeat this process for KNN but instead of α, λ being our tuning parameters it is K. Once we have our best KNN model we continue to the next step.
- Once we have our best KNN and LASSO model we tune them both to the whole training dataset. Then we test them on the test dataset (the 20% data set we havent touched up until this points) and we pick the model (either KNN or LASSO) with the lowest test error.

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)

A ridge regression model performs variable selection by using a penalty term. This will shrink some coefficients to zero. In ordinary least squares, many times our R^2 will increase with more terms making us think the model is fitting better when in fact it is not (it is overfitting 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)

For large values of the tuning parameter λ, our penalty term will be large, thus it will cause the coefficients to shrink to zero / get smaller.

For a tuning parameter near 0, we essentially have the model found by ordinary least squares.

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)

$\hat{y} = 25.293 - 19.780(x_1) - 31.760(x_2) + 2.846(age) - 0.024(age^2) + 2.024 x_1(age) + 2.230 x_2(age) - 0.025 x_1(age^2) - 0.032 x_2(age^2)$

$x_1 = \begin{cases} 1 & \text{married} \\ 0 & \text{otherwise} \end{cases}$ $x_2 = \begin{cases} 1 & \text{never married} \\ 0 & \text{otherwise} \end{cases}$

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

This t-value is useful for determining whether predictors in this model are statistically significant (should be included). Though we should only use this t-value provided for the main effect terms as they do not accurately reflect the significance of the interaction terms

X_1

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)

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

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)

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

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)

It helps model potential non-linear effects of marital status, age, age^2, allowing our model to likely be more flexible.

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: all B_i's are equal to zero

H_A: at least one of the B_i's is not equal to zero

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)

We could have multicollinearity between predictors. It's possible that the inclusion of one or more predictors cause other predictors to appear insignificant, when they could be significant on their own

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)

residual vs. predicted plot