Seat - A39

88 Good job!

ST 563 601 – SPRING 2025 – POST
Exam #1

DELTA

Student's Name: Julia Fish

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, Julia Fish 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 2/6/25 *Date*

FEB 6 8:38AM

Exam must be turned in by: 9:56am *Exam End Time*

Student's Initial Agreement

NOTE: Failure to turn in exam on time may result in penalties at the instructor's discretion.

FEB 6 9:41AM

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)

Predicting chance of parole based on a number of factors regarding the crime, jail time, incidents in jail, etc. (with the goal being to have a generalizable, interpretable model that also has accuracy). model? −1

- Predictive Modeling (4 pts)

Predicting house prices in a neighborhood based on many variables (with the idea being accuracy of prediction over interpretability of model). model? −1

- Pattern Finding (4 pts)

Unsupervised learning, seeing how variables move and if there are relationships to the way that they behave. (Tide of ocean, temperature, humidity, chance of rain, etc. all without trying to use them to predict the other). model? −1

−3

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)

   With the increase in flexibility, we expect a decreased squared bias (negative relationship). That is because, with the model taking more information about the points into account, it is less likely to systematically misrepresent the data.

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

   With the increase in flexibility, we expect an increased variance (positive relationship). That is because, if we were to move a point, the model could change much more than if we did this with a less flexible model.

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

   With the increase in flexibility, we expect a decreased training error. That is because the model can fit the points in the training data better when it is given more flexibility to "move" toward each individual point. (negative relationship).

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

   -2
   "u" shape
   With the increase in flexibility, we expect an increased test error (positive relationship). That is because the flexibility in the model could cause overfitting to the training data, meaning new data will not perform well with this model.

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

   ok
   A tuning parameter/hyperparameter is a parameter that is given a range of different possible values to then select the one that performs the best. This differs from a "regular" parameter due to the "regular" parameters being given one coefficient value (that can change if a different model is fit), and they also represent a variable's connection to the response. Tuning parameters are for selecting the best model, not representing a variable.

   -2

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)

When $p>n$, we can use lasso, ridge regression, elastic net modeling, and PLA. 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.

   False some → 0 -2

   h. LASSO performs shrinkage of coefficient estimates.

   True

-2

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, split the data into training and test set (usually 80/20 or 70/30).

For the lasso model, fit a tuning grid for all potential λ values to be considered is made. Then, split the training set into V-folds for cross validation. For each value of λ, V models will be fit and tested on the left out fold. The model metric will be the average of all V metrics found for this λ value. This is repeated for all values of λ. The most desired model metric (lowest RMSE, etc.) is noted.

For the KNN regression, a tuning grid for all potential K values to be considered is made. Then, split the data into V- folds for cross validation. For each value of K, V models will be fit (leaving one fold out at a time and testing the model with those values). The model metric for that value of K is the average of that metric found on all V- models. This is repeated for all values of K. The model with the most desired model metric (lowest RMSE, etc.) is noted.

The best performing KNN and lasso models (with their specific tuning parameter values) are fit to the entire data set. The one with the more desirable model metric (lowest RMSE, etc.) is the overall best model. -3

-3

8. Consider the Ridge Regression procedure for fitting a multiple linear regression model. With this model we minimize the following criterion (recall $\lambda \geq 0$):

$$\sum_i (Y_i - \beta_0 - X_{i1}\beta_1 - \cdots - X_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

a. What are the benefits of fitting a Ridge Regression model as compared to an ordinary least squares model? (4 pts)

A ridge regression is a good model to use when there is known multicollinearity in the variables. This is because we can shrink the coefficient values appropriately instead of trying to add interactions, dropping variables, or accepting high variances. This model, compared to ordinary least squares, also has more flexibility. -1

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, the coefficient estimates are shrunk more toward 0 (due to the "large" value in front of the $\sum_{j=1}^{p} |\beta_j|$ in the equation).

When the tuning parameter is a value near 0, the coefficients are closer to what they would be for an ordinary least squares regression. That is because the second summation as a whole approaches 0 when its scaling λ value approaches 0, leaving the ordinary least squares model (when $\lambda=0$).

-1

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.

	Without context	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)		25.293	38.116	0.664	0.507
marital_statusmarried	X_1	-19.780	40.405	-0.490	0.624
marital_statusnever_married	X_2	-31.760	40.992	-0.775	0.439
age	X_3	2.846	1.611	1.767	0.077
I(age^2)	X_3^2	-0.024	0.017	-1.470	0.142
marital_statusmarried:age	X_1X_3	2.024	1.716	1.179	0.238
marital_statusnever_married:age	X_2X_3	2.230	1.820	1.225	0.221
marital_statusmarried:I(age^2)	$X_1X_3^2$	-0.025	0.018	-1.412	0.158
marital_statusnever_married:I(age^2)	$X_2X_3^2$	-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)

With context

$$\widehat{wage} = 25.293 - 19.780(married) - 31.760(never\ married) + 2.846(age) - 0.024(age^2) + 2.024(married)(age) + 2.230(never\ married)(age) - 0.025(married)(age^2) - 0.032(never\ married)(age^2)$$

ok

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

The usefulness of this t-value is to represent whether or not that specific variable should be in the model (this t-value is associated with the p-value that is more directly interpretable for this).

0

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)

$\widehat{wage}_{mar,30} = 25.293 - 19.780 + 2.846(30) - 0.024(30^2) + 2.024(1)(30) - 0.025(1)(30^2)$

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)

$\widehat{wage}_{div,30} = 25.293 + 2.846(30) - 0.024(30^2)$

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)

Separate quad models −1

Adding marital status interacting with age (and age^2) allows the impact of age to change based on marital status and vice versa. This allows the model to represent how the variables can interact with one another instead of independently affecting the response at a fixed value of β_j.

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)

$H_0: \beta_1 = \beta_2 = \ldots = \beta_8 = 0$

H_a: at least one $\beta_j \neq 0$ for $j = 1, \ldots, 8$

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 issue could be caused by multicollinearity. If so many variables that are all representing very similar variability in the data (like age & age^2, never married/married with age as well as with age^2, etc.), it will look as if none of the variables themselves explain anything.

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)

For this, we can look at the residual vs. fitted value plot. If this assumption is met, there should be no trumpeting (or pattern at all) to these points.

−1