86

C5

ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: Chris Goodwin

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, Chris Goodwin 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

C Goodwin

STUDENT SIGNATURE

07 Feb 2025

DATE

Exam must be turned in by: 4:05

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 involves classical ideas of statistics like hypothesis testing, confidence intervals, and prediction intervals. Example, simple linear regression to regress SAT on age for example and using F statistic and t stat to see if the predictor is useful to the model.

- Predictive Modeling (4 pts)

This is when we are trying to use our model to make predictions. For example simple linear regression to predict salary based on education. We have a response variable, salary.

ok

- Pattern Finding (4 pts)

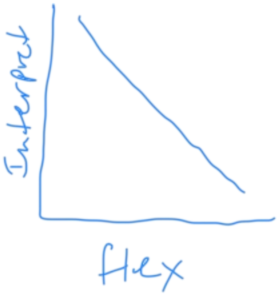
This is like unsupervised learning where we don't have a response variable but are looking for patterns in the data. For example, what might be the relationship between X_1 and X_2

model −2

−2

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 the (squared) bias decreases as the model could match the relationship between the predictors and response better in the training data

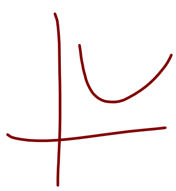
    b. What type of relationship between flexibilty and variance would we expect? Why? (4 pts)

As flexibility increases the variance in the model increases as the data is better able to match the training data.

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

As flexibility increases the training error decreases as the model is able to explain more of the variability in the data.

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



u shape

−2

As flexibility increases, the test error could increase as more flexible models tend not to generalize as well to new data

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

K in KNN is a hyperparameter. It does not relate the predictors to the response as its not part of the model. It just defines the # of observations for which to take an average.

ok

−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)

~~best~~ subset selection
Forward Stepwise
~~backward~~ stepwise
LASSO

−2

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 (if λ large enough)

   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)

1. We want to split the data into training and test sets (eg. 80/20 or 70/30)

2. For Lasso we want to create a tuning grid of L_1 penalty terms (λ).

3. On just the training set we want to split the data into folds with equal n. Within each fold we will split the data using a random process into training and validation sets.

−4

4. We will fit the Lasso model using a λ on the training set of the first fold and then use a metric (MSE) to measure the performance on the validation set. We will do this on each fold and combine the metric for each (eg. average the MSE's)

5. Repeat this process for each λ.

For KNN you do the same except you create a tuning grid of k values (observations for which to take the average).

6. Fit the best LASSO and KNN models on the original ~~test~~ train set —1

compare? —1

7. Fit the best model on the entire data set.

—6

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 ridge regression shrinks the coefficients reducing the variability of their estimates Thereby, decreasing test MSE. (possibly)

multicol. -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 large values the parameters shrink towards 0 but for ridge they don't get there, but shrink to a minimum. If the tuning parameter is near 0, the model is like ordinary least squares

-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.

	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)

ok

$\hat{y} = 25.293 - 19.780(\text{marital status: married}) - 31.760(\text{marital status: never married}) + 2.846(\text{age}) - 0.024(\text{age})^2 + 2.024(\text{marital status: married} \times \text{age}) + 2.230(\text{marital status: never married} \times \text{age}) - 0.025(\text{marital status: married} \times \text{age}^2) - 0.032(\text{marital status: never married} \times \text{age}^2)$

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

Its null hypothesis is that its corresponding $\beta = 0$

$H_0: \beta_j = 0$ v. $H_A: \beta_j \neq 0$. It tells you if that parameter is useful at predicting the response, useful to the model.

ok

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)

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

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)

Separate quadratics

-1

It allows us to have an alternate slope value for how marital statuse changes with age and with age squared.

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)

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

H_a: At least one of the parameters does not equal 0. at least 1 is useful to the model.

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)

multicolinearity in the predictors. You can see this in the st. Error of the predictors being quite large as well.

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)

a residual plot. That is residuals vs fitted values plot.

-1