B25

ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: Koji Takagi

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, Koji Takagi - 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* 2/7/2025

Exam must be turned in by: 2:19 K.T.

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)

using sampling data, we intend to estimate the population mean and or so. For clinical trial, we sample patients and analyze to understand the whole population of the disease.

- Predictive Modeling (4 pts)

kind of supervised, which has response variable. For real world. if we want to know the QOL score in the patients. we can estimate using some parameters such as, age, sex, comorbidity, weight, etc. in the population

- Pattern Finding (4 pts)

Unsupervised doesn't have target, which means no response variable. but want to find the pattern. For example. new biomarkers for cardiovascular disease, we want to find the relation between other parameters.

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 increase, squared bias decrease because Squared brase is average squard difference from the truth.

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

As flexibility increase, variance increase because variance is variability of prediction by themodel. (overfitting)

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

As flexibility increase, training error decrease because. train data is more fitted

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

As flexibility increase, test error first decrease and increase later. like U shape because flexibility cause overfitting.

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

γ in LASSO, k in KNN are different from regular parameter in a parametric model. Non-parametric model need to capture complex model and need to find minimum MSE, which means it

needs to find best model using a tuning parameter.

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, Step forward Selection Step backward selection. elastic model 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.

   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)

- Split train and test dat (70/30, or 80/20)
- LASSO, tuning γ. for example, using 5-fold CV we create 5 groups from the training set. creating 1 test, 4 traing set in the set we learn 4 traing set and perform test set and check MSE. repeated 5 times for each group. calculate average MSE for the first γ. We check the MSE for each γ and find the best γ. using this γ, we perform test model to output MSE.
- Similary, for KNN. tuning k. using 4 training group to find the model and perform test model and repeat 5 times for the first k. same process should be repeated to find the best k.

After that using the k. we perform test model to calculate MSE. Then we compare the MSEs between LASSO and KNN to choose the best model. After choosing the model, we use this model for the entire data.

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)

Less flexible. so. the benefits are as follows:
Low variance, reduce observation needed.

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)

Large λ: Low variance. high Bias, less flexible.

Small λ: close to OLS

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
x_1 marital_statusmarried	-19.780	40.405	-0.490	0.624
x_2 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)

$\hat{y} = 25.293 - 19.78x_1 - 31.76x_2 + 2.846(age) - 0.024(age^2) + 2.024(age \times x_1) + 2.23(age \times x_2) - 0.025(age^2 \times x_1) - 0.032(age^2 \times x_2)$

$x_1 = 1$: married, $x_2 = 1$: never married

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

We check if the |t value| > critical t value or not. the P value can be calculated and found if it is significant or not.

$x_1 = 1$, age $= 30$.

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)

$y = 25.293 - 19.78 + 2.846 \times 30 - 0.024 \times 30^2 + 2.024 \times 30 - 0.025 \times 30^2$.

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)

$x_1 = 0, x_2 = 0$ age $= 30$

$y = 25.293 + 2.846 \times 30 - 0.024 \times 30^2$

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)

provide the other slope of age based on marital_status and of age squared based on 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)

H_0: coefficient $= 0$ vs H_a: coefficient $\neq 0$

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)

Too much parameters cause this issue.

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. fitted plot.