A27

ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: Jarrett Glass

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, Jarrett Glass 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 06 Feb 2025 *Date*

Exam must be turned in by: 5:35 pm JG

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)

- Performing hypothesis testing, such as observing a difference in treatment arms on a clinical trial. (T-tests, etc.)

- Predictive Modeling (4 pts)

- example: looking at sales of homes in a neighborhood, based on # bedrooms and size. what would a 4 bdrm / 2500 sq ft house sell for? (multiple regression model)

- Pattern Finding (4 pts)

- example: reviewing genomic data to look for patterns in genes correlating to disease progression. a model could be generating plots to evaluate colinearity between data points.

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)

Flexibility & bias would be inversely related — a more flexible model could ~~be~~ fit more closely to data points, so it would be less biased against the data.

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

As flexibility increases, so will variance. More fitting in the model increases the susceptibility to noise.

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

A more flexible model would be a better fit to TRAINING data. So increased flexibility implies lower training error.

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

Could not know in advance — increased flexibility could be a benefit, but at a certain point it becomes overfit to training data to negatively impact testing.

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

Tuning parameter helps to optimize a model, but does not give information about the data. A "regular" parameter has something to say about the data.

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)

Ridge regression, LASSO 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.

   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)

kNN: - use createDataPartition(.), generate my train to test split.
Set aside a training set + a test set.
the CV analysis will be done just on the training data set.
when optimum K (based on training MSE) is established,
perform this model using optimum K on just the test dataset.

LASSO: Start similarly. Create separate test/training dataset.
the hyperparameter in this case is "s" (though the slides say t), so that $\sum_{j=1}^{p} |\beta_j| \leq s$.
Perform CV to determine value of s with the lowest MSE, then use Predict( ) to run this on test set.

Compare test MSE of KNN and LASSO methods to determine best model for analysis

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)

It could penalize for particularly large coefficients, and it can be done even in models where p>n like in genomic research, etc.

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)

When λ is large, it will essentially "shut out" our predictors and leave us with a model that only evaluates the intercept.

A λ close to zero is basically the OLS model.

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)

$X_1 = \mathbb{1}\{\text{married}\}$ $X_2 = \mathbb{1}\{\text{never married}\}$ $X_3 = \text{age}$

$\hat{y}$: wage

$$\hat{y} = 25.293 - 19.780X_1 - 31.760X_2 + 2.846X_3 - 0.024X_3^2 + 2.024X_1X_3 + 2.230X_2X_3 - 0.025X_1X_3^2 - 0.032X_2X_3^2$$

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

The t-value, or the p-value derived from the t-value, gives information about whether a specific predictor is significant to the model. In this case no predictor is significant at the 5% level.

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)

$$\text{wage} = 25.293 - 19.780 + (2.846 \cdot 30) - (0.024 \times 30^2) + 2.024(30) - 0.015(30^2)$$

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)

$$\text{wage} = 25.293 + 2.846(30) - 0.024(30^2)$$

(X_1 & X_2 both 0 in this case.)

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)

Having the interaction terms will have an impact on the slopes of the regression lines. $X_3 + X_1X_3$ would be just X_3 if $X_1=0$, or $2X_3$ if $X_1=1$, eg.

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: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$

H_A: Any β_i for $i \in [1, \ldots, 9]$ $\neq 0$.

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)

Could be too much "noise" - possibly no one predictor is having a significant effect. Could perform best subset analysis to determine which predictors are best to use.

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 scatterplot. If points are fairly evenly distributed around a regression line, indicative of even error variance.