79

ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: Dev Kewlani

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, Dev Kewlani 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

Dev Kewlani 02/06/2025

Student Signature *Date*

Exam must be turned in by: 12:09 PM

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.

-2 per part

- Statistical Inference (4 pts)

This is something we ok do when we want to draw inference or interpret the model well. A model with less flexibility and more interpretable helps us do that. Eg. Hypothesis Testing, construction of confidence Inter (Linear regression can be statistically interpreted)

- Predictive Modeling (4 pts)

predictive modelling is usually when we're trying to predict a qualitative/quantitative ok variable and letting the data fit to a response variable. Eg. [Linear Reg, Logistic Reg.

- Pattern Finding (4 pts)

Pattern finding is an unsupervised learning mechanism where we let the data learn ok on its own to find the structure in different parts of it entirely through self-learning. Eg [PCA, k-means, clustering

-6

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)

More flexible ⇒ more parameters ⇒ low squared Bias.
Because it will try to overfit to the training data and capture every data point's relationship w/ it response.

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

More flexible ⇒ (generally) higher variance because it tends to overfit to training data and has high fluctuations on test data, everytime it is tested.

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

More flexible ⇒ lower training error. Again, it fits (tries to) perfectly capture all the points' relationships in training data and tends to overfit.

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

'u' shape -2

Generally, more flexible ⇒ higher test error but it depends on problem to problem. If n is high, it could also (for certain) problems lead to good fit.

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 which can be "tuned" or "optimized" when we are fitting a model on training data. Reg. parameters capture relationship between predictor and output, tuning parameter is something that determine how a model

But generally since overfit on training, performs poorly on test.

-2

may fit onto the data. λ in Lasso, K in KNN are tuning params and determine if we overfit/underfit/.. the data. β's in linear reg. are egs. of reg. param

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)

1) Principal components regression.
2) Forward selection
3) Lasso regression
4) Best subset ($K = 1 \dots n$)
(Elastic Net maybe)

−1

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

−1

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) Split the data into train and test (80%-20%)

On the training data,

KNN

* Initialize a grid of K-values
- For each value of K,
* use a V-fold cross validation (V = 5/10), where you divide the data into 5 parts, 1 part is holdout set, you train the model on 4 parts and test it on the holdout set
* keep a track of test MSEs -

$CV\text{-}test\text{-}K_i = [K_{i1} \dots K_{i5}]$

* Average out the CV-test-Ki & CV-test-λi
* repeat the process for each K
* Select the K-value with least CV-test-Ki

LASSO

Initialize a grid of λ values

For each value of λ

$CV\text{-}test\text{-}\lambda_i = [\lambda_{i1} \dots \lambda_{i5}]$

repeat process for each λ.

Select λ value for least CV-test-λi

* Test on the test set for both k & λ, if test_MSE_k_best < test_MSE_λ_best, fit the model for that k on entire data, else fit the model for that λ on entire data.

model fit on entire training data - 1

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)

* Ridge regression helps in dealing with multi-collinearity issues, ie if predictors that are highly correlated, RR will shrink those coefficients towards 0. If there are large no. of predictors, ridge regr. helps in dealing with this issue. It leads to lower variance than OLS model [slight higher Bias].

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 λ, the coefficients will be shrunk by alot as the loss fn is penalizing [highly] the coefficients, this will make the model underfit the data and will lead to high Bias and high variance.

For λ near 0, RR will behave like an OLS model with no penalization on coeff. estimates and may not generalize to the data well [w.r.t a medium value of λ] so it will have lower Bias but perhaps higher variance.

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)

2 Indicator variables $I_M = \begin{cases} 1 & \text{if married} \\ 0 & \text{if divorced/not married} \end{cases}$

$I_{NM} = \begin{cases} 1 & \text{if not married} \\ 0 & \text{if divorced/married} \end{cases}$

$\hat{y} = 25.293 - 19.78 I_M - 31.76 I_{NM} + 2.846 Age - 0.024 Age^2$
$+ 2.024 (I_M \times Age) + 2.23 (I_{NM} \times Age) - 0.025 (I_M (Age)^2)$
$- 0.032 I_{NM} (Age)^2$

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

t-value helps us in performing hypothesis testing of whether a predictor has any association with a target variable. Based on the t-value, we can get the p-value and we can construct confidence intervals about the range of coefficient estimates. ∎

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)

$\hat{y} = 25.293 - 19.78 + 2.846 \times 30 - (0.024 \times (30)^2) + 2.024 \times 30 - 0.025 \times (30)^2$

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)

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

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)

Adding the interaction term helps us adjust for the slope of the line for a particular age and their marital status in combination. In this example holding age const., a divorced person will have a higher

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.

Separate wage. quadratics −1

i. Write down the null and alternative hypotheses for this global test. (3 pts)

$H_0 \equiv \beta_1 = \beta_2 = \beta_3 \ldots = \beta_j = 0$ [$j = 8$? Don't remember the formula]

$H_a \equiv$ At least one of $\beta_1 \ldots \beta_j$ is not equal to 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)

High Multicollinearity between predictors is making the t-stat shoot very high and causing us to not observe any variable being significant. we should start with only simple terms & see the

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)

effect of individual terms first without including interaction & polynomial terms. −1

Part b: Checking the plot of residuals vs predicted values will help us investigate the homogeneous error variance. The plot should show no discernible pattern and look like white noise.

If the plot looks like "funnel" shape <
it is likely that variance of error terms is not constant and transformations of the response variable or weighted least squares regression may be needed.