

SO

ST 563 601 – SPRING 2025 – POST Exam #1

Student's Name: Constantino Roptis

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, Constantino Roptis
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

Exam must be turned in by:

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)

student can graduate or fail to graduate
we want to find out if a student can graduate or fail to graduate
that will be a classification task with
(graduate/fail to graduate) as the response variable.
model -2

- Predictive Modeling (4 pts)

How much people earn. we will
predict with predictors, such as their age,
education, marital status... Regression task -2
model

- Pattern Finding (4 pts)

Earthquakes, finding patterns on the
stock market (which is extremely difficult) -4

2. Consider having models characterized by flexibility with the scale going from not very flexible to very flexible.

- a. What type of relationship between flexibility and squared bias would we expect? Why? (4 pts)

Not flexible because squared bias is
the model misses truth by squared.
→ 3

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

Flexible since the variance will vary
by the model. ? → 4

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

Flexible because for example overfitting
will present a low training error but ok
underfitting will have a higher training error.

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

Flexible, underfitting will have a higher
test error + can overfitting the model. -4

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

parametric model → fixed structure, few
parameters, low variance
we tune parameters such as the λ in
Ridge regression and k in KNN. -1

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~~, Forward, ~~backward~~ - 2
~~OAS~~, LASSO, Elastic net model

6. State true or false (no need to explain). (3 pts each)

- a. Ordinary least squares performs variable selection.

~~True~~ →

- 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.

~~True~~ →

- e. Ridge Regression performs variable selection.

~~True~~ →

- f. Ridge Regression performs shrinkage of coefficient estimates.

True ✓

- g. LASSO performs variable selection.

~~False~~ →

- 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 test (80/20 or 70/30) ✓
- Tune parameters for each Lasso and kNN
- Using cross validation (5/10) (k closest neighbors)
 - Fit Lasso model
 - test MSE
 - Repeat ^{times} → 3
 - choose the Lasso with the lowest MSE
- Fit kNN
 - Test MSE
 - Repeat for each k
 - choose the k with the lowest MSE
 - choose the k with the lowest MSE → 1
 - Fit both models - to compare
 - on what data?
 - test set
 - choose the best model with the lowest test error
 - fit the overall best model to the entire dataset.

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)

Ridge regression will shrink the coefficients estimates while in OLS model the tuning parameter value is near 0. -2

- 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)

↪ OLS fit

$\uparrow \lambda \rightarrow$ it strictly penalizes the size of the estimate coefficient.

As value near 0 it will be close to the OLS fit.

-2

$$h_1 = \begin{cases} 1 & \text{Married} \\ 0 & \text{Otherwise} \end{cases}$$

$$h_2 = \begin{cases} 1 & \text{Never married} \\ 0 & \text{Otherwise} \end{cases}$$

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)

$$\hat{y} = 25.293 - 19.780h_1 - 31.760h_2 + 2.846(\text{Age}) - 0.024(\text{Age}^2) + 2.024h_1(\text{Age}) + 2.230h_2(\text{Age}) - 0.025h_1(\text{Age}^2) - 0.032h_2(\text{Age}^2)$$

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

We use the t-value to find the p-value.

$P_{\text{t}}(>|t|)$ → we always look the p-value to find if the predictor is statistically significant for our case all p-values are above the 0.05 level, so they are not statistically significant.

std. error is the standard deviation of $\hat{\beta}$ significant

- c. Write down the form of a predicted value for someone that is married and has an age of 30. No need to simplify. (2 pts)

$$\hat{y} = 25.253 - 15.780 + 2.846(30) - 0.024(30^2) + \cancel{2.024(30)} \\ - 0.025(30')$$

- d. Write down the form of a predicted value for someone that is divorced and has an age of 30. No need to simplify. (2 pts)

$$\hat{y} = 25.253 + 2.846(30) - 0.024(30^2)$$

- 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)

c different slope to the model which is good to check if there will be an increase on residual variation.
 Separate quadrates - /

- 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 : the coefficient = 0 \rightarrow we always present the
 H_A : the coefficient $\neq 0$ no effect on H_0 . - 3

- 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) - 3

this means that individual coefficients and interaction are not statistically significant which means that p-value is above 0.05. But we need to take a closer look since the global mode is statistically significant below 0.05

- 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 plot.

resid vs what? - /

Significant
below 0.05