

ST 563 601 – SPRING 2025 – POST

Exam #1

Student's Name: Nick Zehnle

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, Nick Zehnle 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

Nick Zehnle
STUDENT SIGNATURE

2/7/25
DATE

Exam must be turned in by: 12:14

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)

Linear Regression ; interpretable

such that you can infer relationships between predictors and response variable.

Ex: education on salary

- Predictive Modeling (4 pts)

KNN ; non-parametric such that it loses out on interpretability but can fit a training set and produce estimates (prediction) on a test set

Ex: tree height based on tree width

- Pattern Finding (4 pts)

Unsupervised, i.e. no response variable

Ex: patterns in accounting data across firms

(OLS
ridgeless)
parametric

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)

More flexibility \Rightarrow less squared bias.

Better suited to fit more complex relationships.

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

More flexibility \Rightarrow more variance

Overfitting noise.

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

More flexibility \Rightarrow less training error

Can fit more complex relationships
and/or overfit such that training error is very small

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

Unknown. A flexible model may overfit and not generalize well and a non-flexible model may underfit (variance-bias trade-off)

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

A regular parameter is specified in a model while a tuning parameter is found through means such as CV or bootstrapping in the training set.

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 stepwise
- backward stepwise
- LASSO (can set coefficients to 0)

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

a. Ordinary least squares performs variable selection. F

b. Ordinary least squares performs shrinkage of coefficient estimates. F

c. Best subset selection performs variable selection. T

d. Best subset selection performs shrinkage of coefficient estimates. F

e. Ridge Regression performs variable selection. F

f. Ridge Regression performs shrinkage of coefficient estimates. T

g. LASSO performs variable selection. T

h. LASSO performs shrinkage of coefficient estimates. T

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 data into training/test sets
(70/30 or 80/20 for example)
- (2) tune λ and k
 - ↪ use cross-validation (e.g. 5-fold CV)
 - ↪ this creates 5 equal subsets of training set where 4 are selected to train on and 1 to test on
 - ↪ get 5 test MSEs for each λ, k and take average along each λ, k
 - ↪ select λ, k with smallest CV MSE
 - (3) fit models with selected λ, k on entire training set
 - (4) predict on test set and evaluate better test MSE
 - (could use other metrics such as RMSE,
 R^2 , MAE, ...)

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)

By applying a constraint on coefficients, it reduces variability but may introduce bias (i.e. Model becomes less flexible). Useful when OLS is plagued by high variance.

- 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)

$\lambda \rightarrow 0 \Rightarrow RR \rightarrow OLS$

$\lambda \gg 0 \Rightarrow$ coefficients get nearer to 0 but never equal 0
(differs from LASSO in this respect)

$$X_1 = \begin{cases} 1 & \text{married} \\ 0 & \text{o.w.} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{never married} \\ 0 & \text{o.w.} \end{cases}$$

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
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 - X ₃	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)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_3^2 + \hat{\beta}_5 X_1 X_3 \\ + \hat{\beta}_6 X_2 X_3 + \hat{\beta}_7 X_1 X_3^2 + \hat{\beta}_8 X_2 X_3^2$$

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

To assess whether to reject the null of no relationship (i.e. $\hat{\beta}_i = 0$)

- 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)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 + 30\hat{\beta}_3 + 900\hat{\beta}_4 + 30\hat{\beta}_5 + 900\hat{\beta}_7$$

- 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)

$$\hat{Y} = \hat{\beta}_0 + 30\hat{\beta}_3 + 900\hat{\beta}_4$$

- 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)

Produce other slopes $\hat{\beta}_4, \hat{\beta}_5$ for age and age^2 dependent 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)

Null: model 2 doesn't capture more variability of the response than model 1

Alt: reject null

- 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)

Adding more predictors intrinsically reduces variability

- 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)

Resid vs. Fitted



