ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: Kevin Kronk

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, Kevin Kronk 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

Kevin Kronk 2/7/2025

Student Signature *Date*

Exam must be turned in by: 4:10 KK

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)

We want to determine how much a one unit increase in height affects shoe size for our data. Simple linear regression model

$Y = \beta_0 + \beta_1 X_1$ where Y (response) - shoe size, X_1 (predictor) is height.

This is found from our estimate $\hat{\beta}_1$

- Predictive Modeling (4 pts)

We want to predict shoe size based on height.

We can use a simple linear regression model

$Y = \beta_0 + \beta_1 X_1$ where Y (response) is shoe size

X_1 (predictor) is height

with the estimated parameters $\hat{\beta}$ we can make predictions.

- Pattern Finding (4 pts)

Based on torso height and width, clusters the data to determine if there are any patterns, possibly to find a mean torso height and width from 3 groups to make small, medium, and large shirts. We could use heirarchical clustering - each data point starts as its own group, then combine closest ones until we have three groups remaining

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)

A model that is more flexible will have more variance, due to how the observations will vary with a new (test) set of data. This means it will have lower squared bias due to the bias-variance tradeoff. Less flexible model is less able to fit to the shape of the data, therefore, it is more biased (to its fit)

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

The more flexible the model the higher the variance. The more flexible model will be able to fit more closely to the training data. The variability in the data however will cause this fit to be worse on new test data (higher MSE)

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

More flexibility means the model can fit more closely to the training data, therefore training error would be lower.

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

With a closer model fit to the training data, and lower training error, the model wont be able to generalize as well to new testdata, and therefore will have higher test error.

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

We chose a hyperparameter based on a variety of factors like CV error. A regular parameter is found when the model is trained on the data. The hyperparameter determines some aspect of the model used like the number of neighbors (K) in KNN.

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
Forward Selection
Dimensionality Reduction - PCA

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)

- Start by splitting the data into a training set and test set with an 80/20 split.
- Create a grid of λ values for the LASSO model.
- With CV - split the training data into 5 folds (or however many), using 4 to train a model with the first λ value. The 5th fold is used to test the model. Do this 5 times for 5 folds and average the test error (which folds tested and trained on change each time). This is CV error.
- Repeat this process for all the values of λ.
- Then compare and choose the best value of λ, based on the one that minimizes CV error.
- Create a grid of K values for KNN regression.
- Once again for each K use CV in order to find the K that minimizes cross validation error.
- Fit both 'best' models on the full training set.
- Then evaluate them both on the test set. The one with the lower test MSE is the better model.

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)

Because of the penalty term Ridge Regression prevents any B_j coefficient from getting too large (relatively). This shrinks the coefficients overall and makes a model that has lower variance, but more bias. This is important when we have many predictors that give the model too much variance. Even with the increase in bias the decrease in variance leads to lower test MSE.

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 λ the coefficients are heavily penalized and will begin to approach 0, leaving only the intercept, B_0 in the model. For a λ near 0, ridge regression starts turning back into the ordinary least squares solution, where there is no penalty (regularization).

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)

$$\hat{y} = 25.293 - 19.780x_1 - 31.760x_2 + 2.846(age) - 0.024(age^2) + 2.024x_1(age) + 2.230x_2(age) - 0.025x_1(age^2) - 0.032x_2(age^2)$$

$x_1 \begin{cases} 1 & \text{married} \\ 0 & \text{other} \end{cases}$ $x_2 \begin{cases} 1 & \text{never married} \\ 0 & \text{other} \end{cases}$

When x_1 and $x_2 = 0$ we get the baseline, divorced.

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

When testing the hypothesis that $H_0: \beta_j = 0$ $H_A: \beta_j \neq 0$ The test statistic t represents how far away this estimate is from the true β if the true $\beta = 0$ (Null Hypothesis). This can be used to calculate the p-value, how likely this would be to occur under the null hypothesis.

$t = \frac{\hat{\beta}_j - \beta_0}{SE(\hat{\beta}_j)}$ for $H_0: \beta_j = 0$

$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$

$X_1 = 1 \quad X_2 = 0$

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)

$$\hat{y} = 25.293 - 19.780 + 2.846(30) - 0.024(30^2) + 2.024(30) - 0.025(30^2)$$

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) $X_1 = 0 \quad X_2 = 0$

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

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)

It changes the slope of the regression line depending on the 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)

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0 \quad H_A$: At least one $\beta \neq 0$

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)

Multicollinearity in the predictors inflates the estimated standard error of the parameter estimates. Thus the t scores are closer to 0 and the p-values are larger.

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)

The plot of the residuals vs fitted values. If there is a trumpet like shape then there is not constant error variance.