QY Well done!

ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: Henry van Eijk

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, Henry van Eijk 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

HvE 2/7/2025

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)

An example of statistical learning is learning about the relationship between height and weights of men. E.g. a one unit increase in weight causes height to increase by how much. A simple linear regression could be used here and we would be interested in the slope B_1.

- Predictive Modeling (4 pts)

Predictive modeling could be when we want to predict an observation's height given their weight. A KNN model could be used if we only care about prediction, unlike the previous example where we used a very explainable model since the goal was inference.

- Pattern Finding (4 pts)

Pattern finding relates the unsupervised learning where we have no labeled data. Imagine we are given weights only and we want to find some type of relationship between them. We could use some type of clustering unsupervised learning algorithm or an unsupervised learning algorithm for anomaly detection like Isolation Forest.

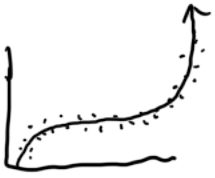
2. Consider having models characterized by flexibility with the scale going from not very flexible to very flexible.

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

As squared bias increases, model flexibility decreases. This is b/c the model becomes more generalized and the variance decreases. E.g., SLR has high bias and is not flexible. In my plot of SLR, the line of predicted values doesn't follow the data very closely. In layman's terms, the SLR can't capture the wiggles of the data.

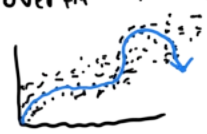
   b. What type of relationship between flexibilty and variance would we expect? Why? (4 pts)

As model flexibility increases, the variance increases and squared bias decreases. The predictions will be a lot more variable since the flexible model can capture a more complicated relationship of the data. In layman's terms, the model with high flexibility/variance can capture the wiggles.



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

As flexibility increases, training error decreases, this is because the model will overfit to the training data. However, it won't perform well on unseen data as it captured too much noise in training set. My plot shows how the model is too flexible with lots of wiggles.



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

The bias/variance tradeoff illustrates the tradeoff between bias (not flexible) and variance (highly flexible). A model with too much flexibility will overfit on the data and have high test error. If a model is not flexible to capture the true relationship of the data, it will also have high test error.

ok

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

In a parametric model, a parameter is something we want to estimate. E.g., μ and σ^2 is normal MLE problem. A hyperparameter doesn't relate to the assumed distribution's parameters. E.g., a hyperparameter could be tuning λ in a Elastic net model. The parameters would be the Beta's $B_0, B_1, \ldots, B_p$ and σ^2. If we assume $y|X \sim N(B_0 + B_1 x_1 + \ldots + B_p x_p, \sigma^2)$, the parameters are specified in the parametric model while λ is not.

ok

0

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)

Elastic net, LASSO, Ridge regression, Best subset selection, Backwards subset selection, Forward subset selection

-2

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 b/c Beta's won't actually =0 but will get close

   f. Ridge Regression performs shrinkage of coefficient estimates.

   True

   g. LASSO performs variable selection.

   True

   h. LASSO performs shrinkage of coefficient estimates.

   True

-2

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)

First, we take our data and split into a training/test set. Possibly 80/20 or 70/20. For Lasso, we need to scale our predictors and create a grid of λ's to tune on. We could use 5 or 10 fold CV for each value of λ. We would pick the λ that gives us the smallest training error (e.g., MSE, MAE). process -3

We could repeat this process for kNN but instead we would tune across the hyperparameter k instead of λ.

Once we have identified the optimal k for kNN and λ for Lasso, we can fit a kNN using all the training data with the optimal k. In addition, we can fit a Lasso across all the training data using the optimal λ.

Once we have our best Lasso and best kNN, we would compute the error for the test set using whichever model metric we used earlier. We would then pick kNN or Lasso depending on which model had smaller testing error.

After picking the best type of model, we could refit across all data using the optimal hyperparameter and we could use this model in real life.

-3

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 is OLS with the L2 penalty term to reduce the magnitude of the Beta coefficients. This introduces bias but will decrease variance. This is helpful b/c it reduces the complexity of our model by having some Betas ≈ 0.

ok

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 $\lambda = 0$, the penalty term = 0 hence OLS. As λ increases the penalty term will cause many of the coefficient estimates to be 0 while this isn't true for small λ since it'll just be OLS.

9

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|)
β_0	(Intercept)	25.293	38.116	0.664	0.507
β_1	marital_statusmarried x_1	-19.780	40.405	-0.490	0.624
β_2	marital_statusnever_married x_2	-31.760	40.992	-0.775	0.439
β_3	age x_3	2.846	1.611	1.767	0.077
β_4	I(age^2) x_3^2	-0.024	0.017	-1.470	0.142
β_5	marital_statusmarried:age $x_1 x_3$	2.024	1.716	1.179	0.238
β_6	marital_statusnever_married:age $x_2 x_3$	2.230	1.820	1.225	0.221
β_7	marital_statusmarried:I(age^2) $x_1 x_3^2$	-0.025	0.018	-1.412	0.158
β_8	marital_statusnever_married:I(age^2) $x_2 x_3^2$	-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) * See my definitions above *

$\hat{y} = 25.3 - 19.8 x_1 - 31.8 x_2 + 2.8 x_3 - 0.02 x_3^2 + 2 x_1 x_3 + 2.2 x_2 x_3 - 0.02 x_1 x_3^2 - 0.03 x_2 x_3^2$

$x_1 = \begin{cases} 1 & \text{married} \\ 0 & \text{o.w.} \end{cases}$

$x_2 = \begin{cases} 1 & \text{never married} \\ 0 & \text{o.w.} \end{cases}$

Divorced is when x_1 and $x_2 = 0$

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

This is a hypothesis test where $\beta_j = 0$ vs. $\beta_j \neq 0$. Assuming $\alpha = 0.05$, this tells us whether that predictor is significant or not. E.g. if p-value > 0.05, the predictor is not useful in our model.

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)

$\hat{y} = 25.3 - 19.8 + 2.8(30) - 0.02(30^2) + 2(30) - 0.02(30^2)$

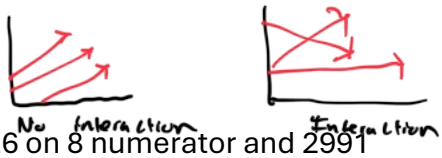
d. Write down the form of a predicted value for somone that is `divorced` and has an age of 30. No need to simplify. (2 pts)

$\hat{y} = 25.3 + 2.8(30) - 0.02(30^2)$

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)

Since the interaction effects include a binary variable, this means the slopes will be different. If we didn't include these the slopes for married, divorced, not married would be parallel.

quadratics here! -1



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)

$H_0: \beta_1, \beta_2, \ldots, \beta_p = 0$ vs. $H_A:$ at least one $\beta_j \neq 0$

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)

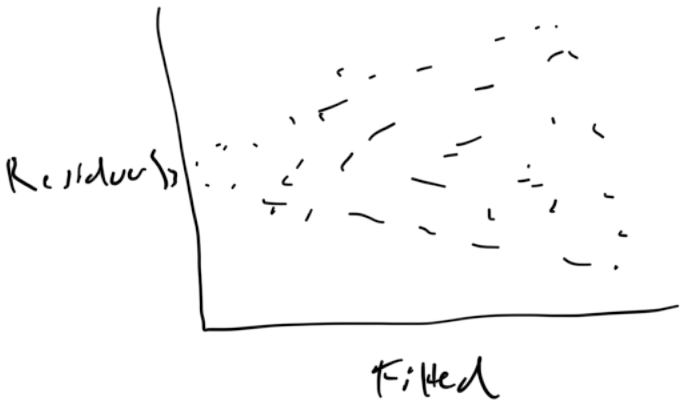
The model is most likely too complicated. We have many predictors and this usually causes multicollinearity. This causes large standard errors causing p-values to be large

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)

-1

A plot with fitted values on X and residuals on y. We check to see if the residuals "fan out"

Non constant variance



Constant Variance