Practice Exam 1 Questions

Your actual exam will have questions similar to those below with space below each question to give answers. I’ve tried to emulate that here but I haven’t taken the time to print the questions out and hand write the answers, which means there may be too much or too little space in this document. I’ll make sure an appropriate amount of space is given on the actual exam (which can also help you guage how detailed of an answer to give!)

* Please note that the example questions below are not exhaustive!
* You may notice there are no programming questions below (that is no R or python syntax at all)
* There are some pseudo code questions. Here you are writing out the logic of the process and how you would go about doing it within a programming language without worrying about the syntax or the process.
* There is very little calculation required for these questions. For most answers that involve output and reading/using it, you do not need to simplify calculations.
* I have not attempted to make the practice exam questions the same length as your actual exam (there will be fewer questions on your actual exam!)
* If you have any other questions about the content or structure of the exam, please post to the discussion forum!

1. When considering the statistical learning paradigm, we discussed supervised and unsupervised learning. What is the major difference between supervised and unsupervised learning?

   
   
   
 

1. When considering predictive modeling as our goal, what is the difference between a regression and classification task?

   
   
   
   
   
 

1. Suppose we collect data to understand how the background of graduate school applicants relates to their graduation status (graduated or failed to graduate). We collect the following information on 200 past students:
   * undergraduate major (Mathematics, Statistics, Other STEM field, Non-STEM field)
   * undergraduate GPA (a continuous 0-4 scale)
   * the time since receiving their undergraduate degree in months
   * whether or not they previously conducted academic research (yes or no).
   1. Would this be a regression or classification task?

* 1. What is the response variable in this scenario?
* 1. Suppose we wanted to turn the undergraduate major predictor into numeric predictors for the purposes of modeling. How many indicator (or dummy) variables are needed to account for this variable? Define these indicator variables.
* 1. Considering your answer to c carefully, what are and in this scenario?

1. Suppose we want to use a multiple linear regression model but the relationships between the predictors and response are non-linear. What are strategies for using a multiple linear regression model to account for these non-linear relationships?



1. Give an example of a non-linear regression model.



1. What is meant by the term ‘parametric model’ vs the term ‘non-parametric model’?

* 1. What are the advantages of using a parametric model over a non-parametric model? What are the disadvantages?
* 1. Give an example of a situation where a non-parametric model would be a preferable choice. The example doesn’t need to be in a real-world context but should be detailed enough to ensure the non-parametric model is a good choice.

1. We discussed the terms overfit and underfit models. Describe these two terms, especially as they relate to bias and variance of the model and training and test errors.
2. 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 kNN model and a ridge regression model. We want to use a train test split and compare the best kNN and ridge regression model on the test set. We wish to determine the appropriate tuning parameters on the training set only using the bootstrap. Fully outline the process for splitting the data, tuning, comparing, and fitting a final overall best model.
3. Describe the curse of dimensionality. Where does this cause problems when doing predictive modeling?



1. When doing predictive modeling, what is a model metric? What is the most commonly used model metric for a regression task?



1. When using the bootstrap for tuning or training our model, what is an out of bag observation and why are they useful?



1. When doing a regression task and considering prediction as our goal, we discussed the expected squared test error given a particular observation,

* We were able to show this could be decomposed into three pieces. Give those three pieces (non-mathematically is fine) and describe what each term represents.

1. Consider the multiple linear regression setting. We discussed using as a way to judge the effectiveness of a model solely on the data on which the model was trained. We saw that use of can be misleading. However, we noted that we could use in the screening step of the best subset selection procedure in order to choose the best model of each size. Fully describe this screening step of the algorithm and explain why the use of here is defensible.
2. Suppose we are fitting a kNN model for a regression task.
   1. Describe how leave-one-out cross-validation (LOOCV) can be used in this setting to choose .

* 1. What are two drawbacks of LOOCV as compared to 5 or 10 fold cross-validation?

1. Consider the LASSO procedure for fitting a multiple linear regression model. With this model we minimize the following criterion (recall ):
   1. What are the benefits of fitting a LASSO model as compared to an ordinary least squares model?

* 1. What happens to our coefficient estimates for a ‘large’ value of the tuning parameter? What happens for a tuning parmeter value near 0?

1. Why is it important to standarize our predictor values when doing a ridge regression, LASSO, or elastic net model?
2. When fitting a regularized or penalized regression model such as the LASSO we discussed using the ‘one-standard error’ method for selecting the tuning parameter. Explain this idea and describe the effect of using this method on the selected model.



1. We perform best subset, forward stepwise, and backward stepwise selection on a single data set with predictors. For each approach we obtain models containing. One model containing 0 predictors, one containing 1 predictor, etc.
   1. Which of the three models with exactly predictors has the smallest training RSS or are we unable to tell?

* 1. Which of the three models with exactly predictors has the smallest test RSS or are we unable to tell?
* + Indicate whether the following statements are true or false
    1. The predictors in the -variable model identified by **forward** stepwise selection are a subset of the predictors in the -variable model identified by **forward** stepwise selection.
    2. The predictors in the -variable model identified by **backward** stepwise selection are a subset of the predictors in the -variable model identified by **backward** stepwise selection.
    3. The predictors in the -variable model identified by **backward** stepwise selection are a subset of the predictors in the -variable model identified by the **best subset** method.
    4. The predictors in the -variable model identified by **best subset** selection are a subset of the predictors in the -variable model identified by the **best subset** method.

1. 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), age, and year that the data was collected. We include an interaction between marital\_status and age in the model. Output for the model is given below.

|  | Estimate | Std. Error | t value | Pr(>|t|) |
| --- | --- | --- | --- | --- |
| (Intercept) | -2526.362 | 722.229 | -3.498 | 0.000 |
| marital\_statusmarried | 26.527 | 11.997 | 2.211 | 0.027 |
| marital\_statusnever\_married | -13.320 | 12.465 | -1.069 | 0.285 |
| age | 0.489 | 0.236 | 2.078 | 0.038 |
| year | 1.299 | 0.360 | 3.608 | 0.000 |
| marital\_statusmarried:age | -0.205 | 0.250 | -0.817 | 0.414 |
| marital\_statusnever\_married:age | 0.321 | 0.280 | 1.146 | 0.252 |

* 1. Write down the fitted equation for . Define any indicator variables as needed.
* 1. One column of the output represents the standard error. What is a standard error generally?
* 1. Write down the form of a predicted value for somone that is married, has an age of 30, and had a year of data collection of 2008. No need to simplify.
* 1. Write down the form of a predicted value for somone that is divorced, has an age of 30, and had a year of data collection of 2008. No need to simplify.
* 1. Consider creating a confidence interval for the mean wage as compared to a prediction interval for a future wage. Which would be wider? Why?
* 1. Conceptually, what does including an interaction between marital\_status and age do to our model as compared to a model without that interaction (that still includes main effects for both, and an effect for year)?
* 1. Based on the above output, which of the **three** predictors would you deem important for the model? State the hypotheses and p-values you used to make this determination.
* 1. Suppose we wanted to check whether the interaction term between marital\_status and age was important for the response, given the other variables in the model. Describe how we could investigate this (you can’t tell with the output above and I’m not expecting you to give the exact formulas - describe the idea of the inferential method we looked at in class for answering this question).
* 1. What type of plot might we look at to investigate the distributional assumption on the errors?
* 1. A few observed values and predicted values are given in the table below. What is the residual for the first observation. No need to simplify.

| wage | predicted |
| --- | --- |
| 75.04315 | 81.10693 |
| 70.47602 | 83.36921 |
| 130.98218 | 115.28895 |