Final study guide:

Final Wed May 10th at 8:30 am (starting at 8:30 and not 8:00; exam is scheduled till 10:30)

Review sessions in lab ( Wed 26 ) and we will hold a review session in our last lecture (Monday May 1)

Questions on the final will be a mix of questions on this study guide and questions you haven’t seen before.

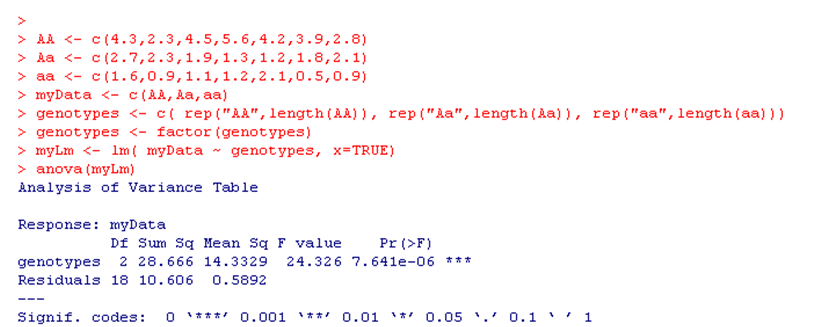
1. You are responsible for the material on the mid-term study guide:

http://afodor.github.io/classes/stats2017/MidtermStudyGuide.docx

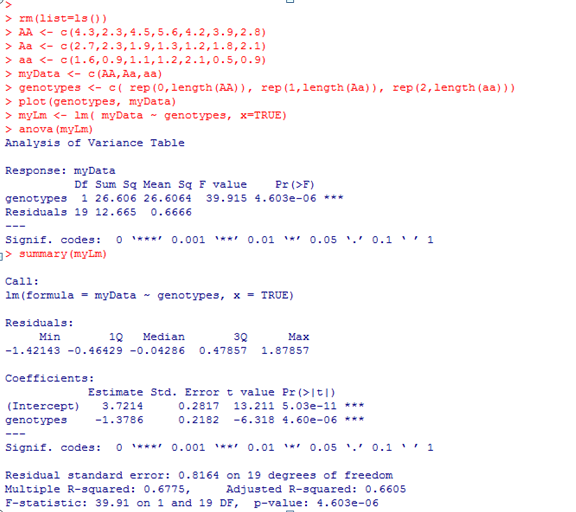
In general, you are responsible for the material in the lectures (on the PowerPoint slides).

Finally, you are responsible for the material in the labs. If you were asked to perform an operation in a lab exercise, you may be asked to perform a similar operation on the final.

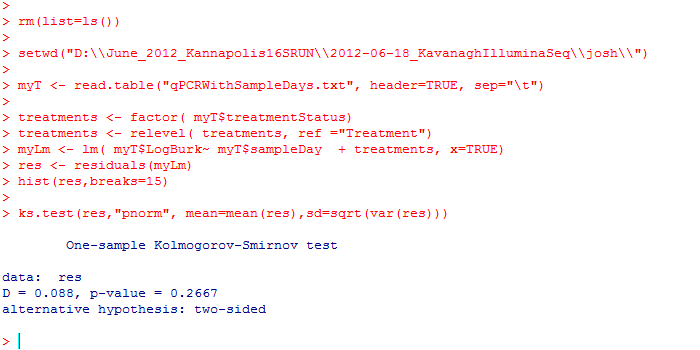
(2) The following is typed into R. What is the null hypothesis being evaluated by the p-values? What assumptions went into producing that p-value? Write the equations for the full and reduced model. Draw graphs representing the full and reduced model.



(3) The following is typed into R. What are the null hypotheses being evaluated by the p-value in the call to summary? What assumptions went into producing that p-value? Write the equations for the full and reduced models. Draw graphs representing the full and reduced model.

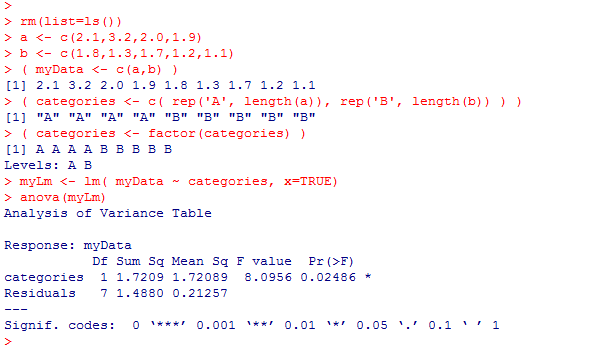


(4) Given the below results, does the assumption of normality seem appropriate for this linear model? Justify your answer.

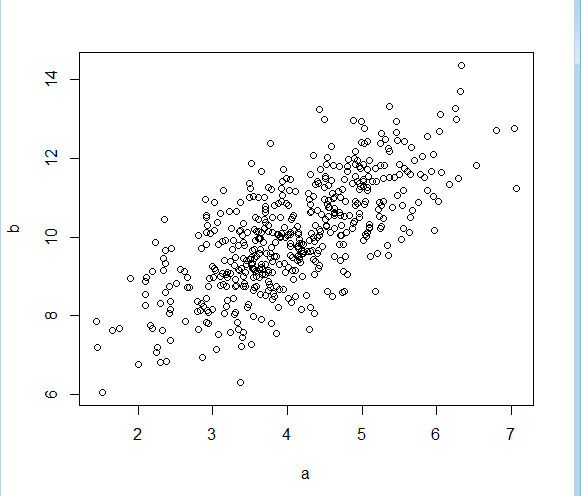


|  |  |
| --- | --- |
|  |  |
|  |  |

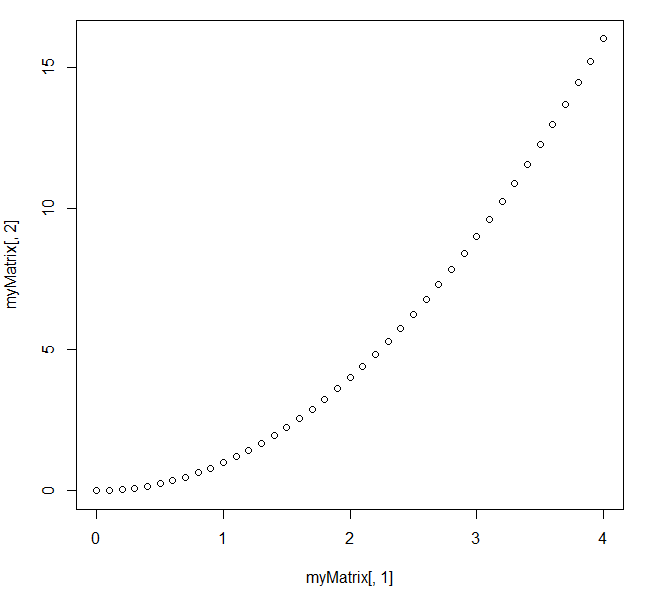
(5) The following is typed into R. What is the null hypothesis being tested. What are the assumptions being used to generate the p-values. Is this a one-sided or two-sided test? What is an alternative test that could relax the assumption of equal variance?

****

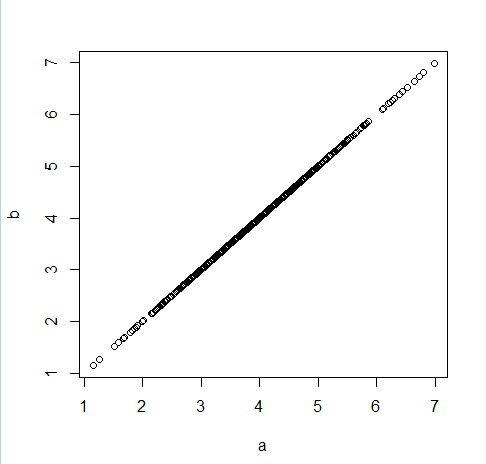
(6) An analyst tells you that “a” and “b” below are the first two principle components of a multi-dimensional dataset? Is this possible? Why or why not?



(7) Draw a graph representing the below data transformed into the first two principle components.



(8) In a PCA transformation of the below data, what % variance would be explained by the first principle component? What % variance would be explained by the 2nd principle component?



(9) How is the AIC criteria defined? When is a model “better” under the AIC criteria? In a maximum likelihood fit, why not always just take the most likely model?

(10) You have an experimental design in which 10 hospitals use drug A and 10 other hospitals use drug B. You measure the weight of each patient in the study. What are some of the limitations of a simple lineal model like:

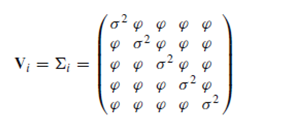
lm( weight ~ drug + hospital)

to analyze your data? How can you analyze your data to test the null hypothesis that drug has no effect?

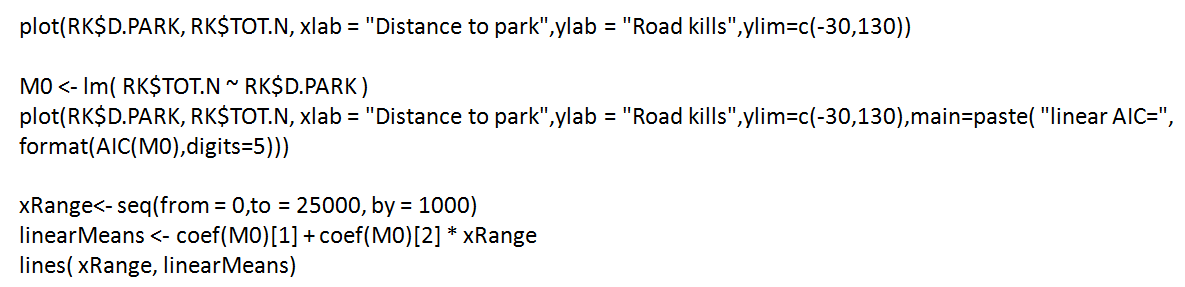
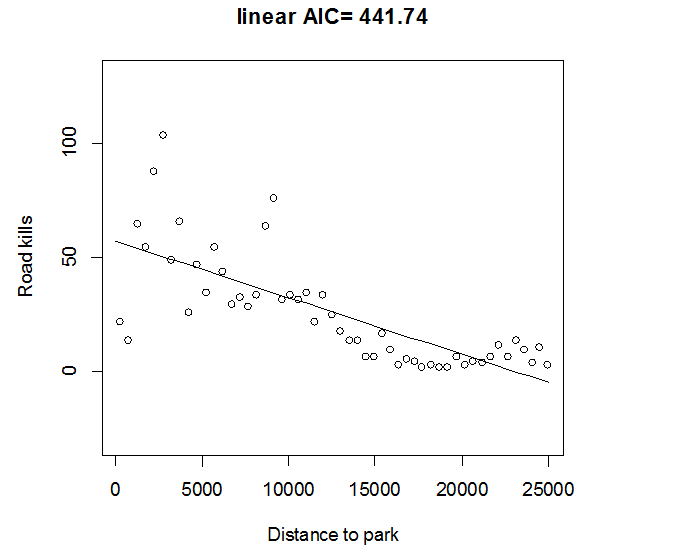
(11) What are the different assumptions of a variance-covariance matrix for residual errors in a linear model that looks like this:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **σ2** | **0** | **0** | **0** | **0** |
| 0 | σ2 | 0 | 0 | 0 |
| 0 | 0 | σ2 | 0 | 0 |
| 0 | 0 | 0 | σ2 | 0 |
| 0 | 0 | 0 | 0 | σ2 |

Vs. a variance-covariance matrix for residuals that looks like this:



(12) What are two potential problems with the model shown below? What are some alternative models that might solve these problems?

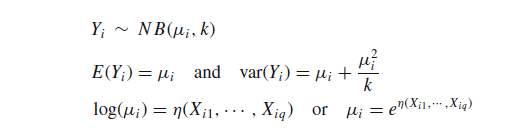


(13) What is the relationship between a “standard” linear model and generalized linear models based on the Poisson and negative binomial distributions? When is it appropriate to use one over the other?

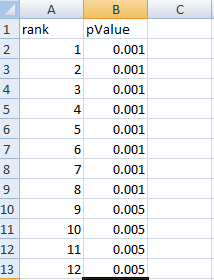
(14) What is the relationship between a logistic regression and the binomial distribution? When should you use a logistic regression?

(15) What is a zero-inflated generalized linear model?

(16) Understand how these equations define a generalized linear model based on the Negative Binomial distribution (and the equivalent equations for the Poisson distribution and for logistic regression)



(17) Consider a microarray experiment with 1,000 genes. An experiment that tests two conditions (e.g. cancer vs. non-cancer) reports the following for its 10 most significant genes…

Using the Benjamini and Hochberg method, how many genes would be significant at a 50% false discovery rate threshold?

Using Bonferroni correction what would be the p-value threshold required for a family-wise error rate of 0.05?

(18) What is the difference between the forward and backwards algorithm. Understand how to apply the two algorithms for posterior decoding (as in the final lab exercise).

(19) What is the difference between the Viterbi algorithm and the forwards algorithm.

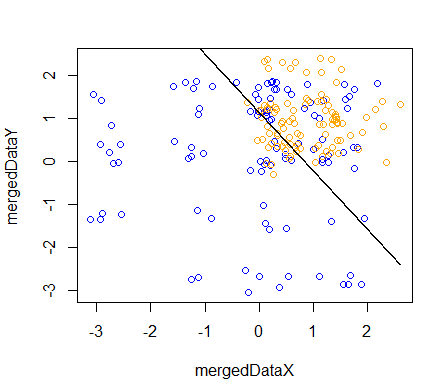
(30) What does the Baum-Welch algorithm do? What does the sentence “the Baum-Welch algorithm can get stuck in local minima” mean.

(31) What does the sentence: “First order Markov chains have no memory” mean?

(32) What is the difference between emission and transmission probabilities in a Markov chain?

(33) When working with posterior decoding on long strings, why do we need to work in log(probabilities)?

(34) In the below dataset, why would a k-means nearest neighbor classification model yield better results than the linear discrimination model below?



(35) In the above graph, a nearest neighbor model with k=1 could correctly classify every data point as blue or orange? Why not just always use that model as a classifier? What is the bias vs. variance trade-off?

(36) As the number of nearest neighbors goes up, does the complexity of the model go up or down? Is a complex model more likely to underfit or overfit a training data set?

(37) How do bias vs. variance trade off impact parameter selection in support vector machines?

(38) Draw an ROC curve from a model that can perfectly separate two conditions. Draw an ROC curve from a model that guesses two conditions at random. Do either of these cases describe the ROC curve below? How good is the classification scheme shown below?

