Final Exam

Unless otherwise specified, assume all α (p-value) thresholds to be 0.05, and all tests to be two-sided if that is an option. All calculations may be done with R or by hand unless otherwise specified; when a problem calls for something "by hand," feel free to use R as a calculator for basic operations like sums and means, and please show and explain your work as much as possible, using latex for displaying math whenever that would make it clearer. All problems are worth equal amounts except problems 1, 6a, 7a, and 8a, which are worth twice as much.

Some hints for creating the simulated data: To create 100 observations where $y = 1 + 10x + \epsilon$ and x and ϵ have mean 0 and sd 1, you can do `x <- rnorm(100)` and `y <- 1 + 10*x + rnorm(100)`. To create y as a function of 10 x variables, all normal with mean 0 and sd 1, where the coefficient on the first 5 variables is 1 and the second is 10 (ie, y = x1 + x2 + ... + 10x5 + 10x6 + ...), try `xmat <- matrix(rnorm(100*10),100,10)` and then use matrix multiplication to get y: `y <- xmat %*% c(rep(1,5),rep(10,5)) + rnorm(100)`. And to create a factor that is "a" when x < 0 and "b" when x >= 0, a useful function is `cut`, eg `myfactorvar <- cut(x,breaks=c(-Inf,0,Inf),labels=c("a","b"))`; see `?cut()` for more, though you can also just do it manually via a loop as in the lesson.

Oh, and please use `set.seed(1)` somewhere at the beginning of your RMarkdown so that your simulated data matches everyone else's. Note that this only has to be done once, not for each separate code block.

Good luck!

1. You roll five six-sided dice. Write a script in R to calculate the probability of getting between 15 and 20 (inclusive) as the total amount of your roll (ie, the sum when you add up what is showing on all five dice). Exact solutions are preferable but approximate solutions are ok as long as they are precise.

2. Create a simulated dataset of 100 observations, where x is a random normal variable with mean 0 and standard deviation 1, and $y = 0.1 + 2 * x + \epsilon$, where epsilon is also a random normal error with mean 0 and sd 1.

a. Perform a t test for whether the mean of Y equals the mean of X using R.

b. Now perform this test by hand using just the first 5 observations. Please write out all your steps carefully.

c. Assuming the mean and sd of the sample that you calculated from the first five observations would not change, what is the minimum total number of additional observations you would need to be able to conclude that the true mean μ of the population is different from 0 at the p = 0.01 confidence level?

3. Generate a new 100-observation dataset as before, except now $y = 0.1 + 0.2 * x + \epsilon$

a. Regress y on x using R, and report the results. Discuss the coefficient on x and its standard error, and present its 95% CI.

b. Use R to calculate the p-value on the coefficient on x from the t statistic for that coefficient as shown in the regression in 3a, and confirm that your p-value matches what is shown in 3a. What does this p-value represent (be very precise in your language here)?

c. Use R to calculate the p-value associated with the F statistic reported in your regression output. What does this test and its p-value indicate?

d. Using just the first five observations from your simulated dataset, calculate by hand the coefficient on x, its standard error, and the *adjusted* R^2. Be sure to show your work, but you may use R for the simple math.

4. Now generate $y = 0.1 + 0.2 * x - 0.5 * x^2 + \epsilon$ with 100 observations.

a. Regress y on x and x^2 and report the results. If x or x^2 are not statistically significant, suggest why.

b. Based on the known coefficients that we used to create y, what is the exact effect on y of increasing x by 1 unit from 1 to 2?

c. Based on the coefficients estimated from 4(a), what is the effect on y of changing x from -0.5 to -0.7?

5. Now generate x_2 as a random normal variable with a mean of -1 and a sd of 1. Create a new dataset where $y = 0.1 + 0.2 * x - 0.5 * x * x_2 + \epsilon$.

a. Based on the known coefficients, what is the exact effect of increasing x_2 from 0 to 1 with x held at its mean?

b. Regress y on x, x_2, and their interaction. Based on the regression-estimated coefficients, what is the effect on y of shifting x from -0.5 to -0.7 with x_2 held at 1?

c. Regress y on x alone. Using the R^2 from this regression and the R^2 from 5(b), perform by hand an F test of the complete model (5b) against the reduced, bivariate model. What does this test tell you?

6. Generate a dataset with 300 observations and three variables: f, x1, and x2. f should be a factor with three levels, where level 1 corresponds to observations 1-100, level 2 to 101-200, and level 3 to 201-300. (Eg, f can be "a" for the first 100 observations, "b" for the second 100, and "c" for the third 100.) Create x1 such that the first 100 observations have a mean of 1 and sd of 2; the second 100 have a mean of 0 and sd of 1; and the third 100 have a mean of 1 and sd of 0.5. Create x2 such that the first 100 observations have a mean of 1 and sd of 2; the second 100 have a mean of 1 and sd of 1; and the third 100 have a mean of 0 and sd of 0.5. (Hint: It is probably easiest to create three 100-observation datasets first, and then stack them with rbind(). And make sure to convert f to a factor before proceeding.)

a. Using the k-means algorithm, peform a cluster analysis of these data using a k of 3 (use only x1 and x2 in your calculations; use f only to verify your results). Comparing your clusters with f, how many datapoints are correctly classified into the correct cluster? How similar are the centroids from your analysis to the true centers?

7. Generate a dataset of 200 observations, this time with 90 independent variables, each of mean 0 and sd 1. Create y such that:

$$y = 2x_1 + ... + 2x_{30} - x_{31} - ... - x_{60} + 0 * x_{61} + ... + 0 * x_{90} + \epsilon$$

where ϵ is a random normal variable with mean 0 and sd 10. (Ie, the first 30 x's have a coefficient of 2; the next 30 have a coefficient of -1; and the last 30 have a coefficient of 0.)

a. Perform an elastic net regression of y on all the x variables using just the first 100 observations. Use 10-fold cross-validation to find the best value of λ and approximately the best value of α.

b. How accurate are your coefficients from (a)? Summarize your results any way you like, but please don't give us the raw coefficients from 90 variables.

c. Using the results from (b), predict y for the second 100 observations. How accurate is the MSE of your prediction?

d. Compare the MSE of (c) to the out-of-sample MSE using ordinary multiple regression. Explain your results, including if the regular regression failed for any reason.

8. Use the data from 7 to generate a new y2 that is 1 if $y > 0$ and 0 otherwise.

a. Using the same process as in 8, estimate an SVM model of y2 on all the x variables for the first 100 variables. Use 10-fold cross-validation to select the best kernel.

b. Using the fitted model from (a), predict y2 for the second 100 observations, and report the accuracy of your predicted y2.