Exercises 9 · Portfolio modeling; logistic regression

Problem 1) bivariate normal model for stocks and bonds

Complete the case study from class on the bivariate normal model for stocks and bonds: `https://github.com/jgscott/learnR/blob/master/cases/bvnorm/bvnorm.md`. You don't need to do the part marked as optional. For this one, all you need to do is quote numbers to answer the questions, and to provide histograms like those in Figure 15.9 of the course packet.

Problem 2) bootstrap resampling in portfolio modeling

Suppose you have $10000 to invest, and are contemplating allocating your wealth among five major asset classes, as represented by five exchange-traded funds:

- SPY, large-cap stocks
- TLT, long-term government bonds
- LQD, long-term corporate bonds
- DBC, commodities (oil, gold, metals, etc)
- VNQ, real estate

You will use Monte Carlo simulation combined with bootstrap resampling to compare the short-term risk profiles for the following three portfolio allocations:

1. 50% of your wealth in stocks (SPY), 50% in real estate (VNQ)
2. 50% in corporate bonds (LQD), 50% in government bonds (TLT)
3. 20% in each of the five asset classes.

Daily data for all five funds is available using the `quantmod` package; see the `portfolio.R` script on the class website.

Suppose you plan to take a buy-and-hold strategy for four weeks (20 trading days). That is, you do not rebalance your portfolio along the way to compensate for differing gains and losses in each of the assets. Briefly acquaint yourself (using outside sources) with the concept of *value at risk* (VaR) for a financial portfolio. The Wikipedia entry on VaR (clickable link here) should be sufficient for this purpose, but there are plenty of other web resources as well.

Your task: use Monte Carlo simulation coupled with bootstrap resampling to estimate the 5% Value at Risk (VaR) for each of these portfolios over this horizon. In light of these numbers, which portfolio looks riskiest, and which least risky?

To get enough Monte Carlo draws, you may need to leave your computer running awhile in the background. For the purpose of debugging, I would recommend checking your scripts first with relatively small Monte Carlo sample sizes (e.g. 10), just to make sure they work as intended. Then increase the Monte Carlo sample size and take a break while your computer churns through the computations necessary to get your final answer.

Problem 3) A hospital audit: assessing radiologist performance

The case

The data in "brca.csv" consist of 987 screening mammograms[1] administered at a hospital in Seattle, Washington. Five radiologists, each of whom frequently read mammograms, were selected at random from those at the hospital. For each radiologist, roughly 200 of the mammograms each had read were selected at random. Each row of the data set corresponds to a single woman's mammogram. The radiologist who read it is identified by a three-number code (1-999).

[1] https://www.nlm.nih.gov/medlineplus/mammography.html

For each patient, two outcomes are recorded. The first is an indicator of whether the patient was recalled by the radiologist for further diagnostic screening after the radiologist read the mammogram (1=Recalled for further diagnostic screening, 0=Not recalled). The second outcome is an indicator of whether there was an actual diagnosis of breast cancer within 12 months following the screening mammogram (1=Yes, 0=No). Ideally, the radiologist should be: (1) minimizing false negatives, i.e. recalling the patients who do end up getting cancer, so that they can be treated as early as possible; and (2) also minimizing false positives, i.e. not recalling the patients who do not end up getting cancer, so that they are not alarmed unnecessarily. Of course, this ideal not attainable. Mammography is inexact, and sometimes there will be mistakes.

The data

In addition to the cancer and recall outcomes, several risk factors for breast cancer identified in previous studies are provided in the data set. Referent values for a "typical patient" are indicated by asterisks:

age: 40-49*, 50-59, 60-69, 70 and older
family history of breast cancer: 0=No*, 1=Yes
breast cancer symptoms: 0=No*, 1=Yes
menopause/hormone-therapy status: Pre-menopausal, Post-menopausal & no hormone replacement therapy (HT), Post-menopausal & HT*, Post-menopausal & unknown HT
breast density classification: 1=Almost entirely fatty, 2=Scattered fibroglandular tissue*, 3=Heterogenously dense, 4=Extremely dense

Audit goals

The goal of this case study is to examine the performance of the radiologists. This kind of statistical audit is a crucial link in the chain of modern evidence-based hospital practice. Specifically, your audit should address two questions.

(A) Are some radiologists more clinically conservative than others in recalling patients, holding patient risk factors equal?

Some advice: imagine two radiologists who see the mammogram of a single patient, who has a specific set of risk factors. If radiologist A has a higher probability of recalling that patient than radiologist B, we'd say that radiologist A is more conservative (because they have a lower threshold for wanting to double-check the patient's results). So if all five radiologists saw the same set of patients, we'd easily find out whether some radiologists are more conservative than others.

The problem is that the radiologists don't see the same patients. So we can't just look at raw recall rates—some radiologists might have seen patients whose clinical situation mandated more conservatism in the first place. Can you build a regression model that addresses this problem, i.e. that holds risk factors constant in assessing whether some radiologists are more conservative than others in recalling patients?

(B) When the radiologists at this hospital interpret a mammogram to make a decision on whether to recall the patient, does the data suggest that they should be weighing some clinical risk factors more heavily than they currently are?

Again, some advice: let's focus on family history as a specific risk factor (a similar line of reasoning applies to any risk factor). Consider two different regression models: Model A, which regresses a patient's cancer outcome on the radiologist's recall decision; and Model B, which regresses a patient's cancer outcome on the radiologist's recall decision AND the patient's family history. Now imagine that the radiologist was appropriately accounting for a patient's family history of breast cancer in interpreting the mammogram and deciding whether to recall the patient for further screening. If that were true, would you expect that Model B would be any better than Model A at predicting cancer status? Why or why not? If instead it turns out that Model B is significantly better than Model A, what does that say about the radiologist's process for making a recall decision?