Summary Notes: Part 2

James G. Scott

Spring 2015

This set of summary notes is not exhaustive, but is a good place to start. Just because something fails to appear here does not mean you are not responsible for knowing it. You'll also recognize that this summary heavily emphasizes the "what" and not so much the "why." For the "why," you'll need to consult the course packet, the notes, and the homeworks.

1 Model choice

Analysis of variance. Key facts about an analysis of variance (ANOVA):

- An ANOVA table partitions the predictable variation in a regression model among the different variables. By predictable variation, we mean the PV in the variance decomposition $\text{PV} = \text{TV} + \text{UV}$ that is used to define R^2.
- The table is constructed sequentially, adding one variable at a time and tracking the change in predictable variation.
- An ANOVA is therefore order dependent when the variables are correlated (collinear). In other words, the attribution of "credit" to each individual predictor is ambiguous.

Occams's razor. Key ideas:

- Occam's razor is the philosophical principle that explanations (models) should be only as a complex as they need to be in order to explain reality well.
- We've refered to this as the fit/simplicity tradeoff: adding more variables increases the fit and decreases the simplicity.
- Permutation tests and F test can be used to address focused questions about whether adding an individual variable to the model is warranted. See course packet for details.
- AIC (Akaike information criterion) can be used as a quick measure to compare different models in terms of which is optimally balancing fit and simplicity.
- AIC is useful for pure-prediction problems, where we don't care too much about interpreting the model.

- Stepwise selection (in which variables are added/deleted sequentially) is a way to attempt to find a model that has the best AIC from among a large candidate set of models.

2 Logistic regression

The linear probability model. In many situations, we would like to forecast the outcome of a binary event, given some relevant information:

- Given the pattern of word usage and punctuation in an e-mail, is it likely to be spam?
- Given the temperature and cloud cover on Christmas Eve, is it likely to snow on Christmas?
- Given a person's credit history, is he or she likely to default on a mortgage?

In all of these cases, the y variable is the answer to a yes-or-no question. Nonetheless, we can still use regression for these problems. Let's suppose, for simplicity's sake, that we have only one predictor x, and that we let $y_i = 1$ for a "yes" and $y_i = 0$ for a "no." One naïve way of forecasting y is simply to plunge ahead with the basic, one-variable regression equation:

$$\mathrm{E}(y_i \mid x_i) = \beta_0 + \beta_1 x_i \,.$$

Since y_i can only take the values 0 or 1, the expected value of y_i is simply a weighted average of these two cases:

$$\begin{aligned} \mathrm{E}(y_i \mid x_i) &= 1 \cdot P(y_i = 1 \mid x_i) + 0 \cdot P(y_i = 0 \mid x_i) \\ &= P(y_i = 1 \mid x_i) \end{aligned}$$

Therefore, the regression equation is just a linear model for the conditional probability that $y_i = 1$, given the predictor x_i:

$$P(y_i = 1 \mid x_i) = \beta_0 + \beta_1 x_i \,.$$

This model allows us to plug in some value of x_i and read off the forecasted probability of a "yes" answer to whatever yes-or-no question is being posed. It is often called the linear probability model, since the probability of a "yes" varies linearly with x.

The logistic link function. The linear probability model is perfectly reasonable in many situations. But suffers from a noticeable problem. The left-hand side of the regression equation, $P(y_i = 1 \mid x_i)$, must be between 0 and 1. But the right-hand side, $\beta_0 + \beta_1 x_i$, can be any real number between $-\infty$ and ∞. We'd be better off with some transformation g that takes an unconstrained number from the right-hand side, and maps it to a constrained number on the left-hand side:

$$P(y_i \mid x_i) = g(\beta_0 + \beta_1 x_i) \,.$$

Such a function g is called a *link function*. A model that incorporates such a link function is called a *generalized linear model*; and the part inside the parentheses $(\beta_0+\beta_1 x_i)$ is called the *linear predictor*, and is often denoted as ψ_i.

We use link functions and generalized linear models in most situations where we are trying to predict a number that is, for whatever reason, constrained. Here, we're dealing with probabilities, which are constrained to be no smaller than 0 and no larger than 1. Therefore, the function g must map real numbers on $(-\infty,\infty)$ to numbers on $(0,1)$. It must therefore be shaped a bit like a flattened letter "S," approaching zero for large negative values of ψ_i, and approaching 1 for large positive values.

With multiple regressors $(x_{i1},\ldots,x_{ip})$, we have

$$\Pr(y_i=1\mid x_i)=w_i=\frac{\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})}{1+\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})}. \tag{1}$$

Recall that odds are just a different way of expressing probabilities:

$$(\text{Odds that } y_i \text{ is } 1)=O_i=\frac{w_i}{1-w_i}.$$

If you churn through the algebra and re-express the logistic-regression equation (1) in terms of odds, you will see that the log-odds of success—or equivalently the *logit transform* of the success probability—are being modeled as a linear function of the predictors:

$$\operatorname{logit}(w_i)=\log O_i=\log\left(\frac{w_i}{1-w_i}\right)=\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}.$$

3 Time Series

Key ideas:

- A time series $y_t : t=1,\ldots,N$ is a set of observations collected sequentially in time. The subscript t may refer to any time unit, e.g. seconds, days, weeks, years.
- To model a linear trend in a time series, introduce a time index and regression on this index:

$$y_t=\beta_0+\beta_1 t+\epsilon_t.$$

- To model seasonal variation, or periodic variation more generally, introduce dummy variables corresponding to the "seasons" (months, quarters, weeks, etc).

4 Probability

Three common interpretations of probability:

1. The axiomatic interpretation.
2. The frequentist ("Vegas") interpretation.
3. The Bayesian/subjectivist ("Wall Street") interpretation.

Kolmogorov's axioms. The following are called Kolmogorov's axioms, after the Russian mathematician of the same name.

1. $0 \leq P(A) \leq 1$ for any event A. In words: all probabilities are numbers between zero and one.
2. If Ω is a certainty, then $P(\Omega) = 1$. In words: events that are certain to occur have probability one.
3. If A and B cannot both happen, then $P(A \text{ or } B) = P(A) + P(B)$. In words: the probabilities for mutually exclusive events add together.

These three axioms can be shown to be a consequence of rational behavior via the Dutch-book argument outlined in the notes.

Odds. The following simple conversion formulas are useful:

$$\text{Odds Against A} = \frac{1 - P(A)}{P(A)} \tag{2}$$

$$P(A) = \frac{1}{(\text{Odds Against A}) + 1}, \tag{3}$$

where "Odds Against A" is interpreted as a decimal number (e.g. odds of 9:2 are $9/2 = 4.5$).

Important rules for probability. Here are several important probability rules. I will provide a list of these on the exam.

Addition rule: The probability that either A or B will happen is

$$P(A \cup B) = P(A) + P(B) - P(A, B), \tag{4}$$

where $P(A, B)$ is the probability that both A and B happen at once.

Multiplication rule: The joint probability that A and B will both happen is

$$P(A, B) = P(A) \cdot P(B \mid A), \tag{5}$$

where $P(B \mid A)$ is the conditional probability that B will happen, given that A happens.

Bayes' rule: a rule for updating prior probabilities into posterior probabilities, given new data.

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}.$$

Notice how the addition rule, which works for any events A and B, differs from Kolmogorov's third axiom, which makes the restrictive assumption that A and B are mutually exclusive. If A and B cannot both occur, then $P(A,B) = 0$ and we're back to the original rule.

5 Random variables

A random variable is anything we're not sure about. We describe our uncertainty about a random variable X using a probability distribution $P(X)$.

If two random variables X and Y are coupled (not independent), then we describe them with a joint distribution $P(X,Y)$. A joint is an exhaustive description of joint outcomes for two or more variables at once, together with the probabilities for each of these outcomes. For example, the table below depicts a simple, stylized joint distribution for the rain and average wind speed on a random day in February.

Outcome	Wind (mph)	Rain (inches)	Probability
1	5	1	0.4
2	5	3	0.1
3	15	1	0.1
4	15	3	0.4

Moments. Moments are summaries of a probability distribution. Three examples are important moments are the expected value, the variance, and the covariance.

The *expected value* of a probability distribution is a probability-weighted average of the possible values of the random variable. If the random variable has N possible outcomes $\{x_1, \ldots, x_N\}$ having corresponding probabilities $\{p_1, \ldots, p_N\}$, then the expected value is

$$E(X) = \sum_{i=1}^{N} p_i x_i .$$

Next, the *variance* of a probability distribution is a measure of dispersion. It is "the expected squared deviation from the expected value", or

$$\text{var}(X) = \text{E}\left(\{X - \text{E}(X)\}^2\right).$$

Finally, covariance measures how two variables in a joint distribution are coupled.

$$\text{cov}(X,Y) = E\Big\{[X - E(X)][Y - E(Y)]\Big\} = \sum_{i=1}^{n} p_i \left[x_i - E(X)\right]\left[y_i - E(Y)\right].$$

Functions of random variables. We often construct a new random variable as a function of other random variables, like when forming a portfolio of financial assets. The simplest example is a linear combination:

$$W = aX + bY + c$$

for some random variables X and Y and some constants a, b, and c.

The moments of the linear combination W can be described in terms of the moments of X and Y.

$$E(W) = aE(X) + bE(Y) + c \tag{6}$$

$$\text{var}(W) = a^2 \text{ var}(X) + b^2 \text{ var}(Y) + 2ab \text{ cov}(X,Y). \tag{7}$$

For nonlinear functions of random variables, there are no simple formulas for the variance or expected value. In fact, Jensen's inequality says that if $W = f(X)$ and f is a nonlinear function, then $E(f(X)) \neq f(E(X))$. Thus we often turn to simulation as a way of approximating the moments of difficult functions of random variables.

6 Utility

Your utility function is something that encodes your preferences between different bets or lotteries. We'll use the notation "$P \succ Q$" to express mathematically the idea that you'd rather play lottery P than Q. The binary relation "$\succ$" is called your *preference relation*.

The *von Neumann–Morgenstern rules* are some very simple, common-sense constraints on your preferences. The rules are:

(1) Your preferences are complete. That is, for any two lotteries P and Q, you are able to state whether $P \succ Q$, $Q \succ P$, or $P \sim Q$ (the third option indicating your indifference between the two.) You're not allowed to say, "I don't know."

(2) Your preferences are transitive. That is, if $P \succ Q$ and $Q \succ R$, then $P \succ R$.

(3) You are willing to "split the difference" between favorable and unfavorable options. This one is often called the rule of continuity. Suppose there are three lotteries P, and Q, and R, such that P is your favorite choice, R is your least favorite, and Q is somewhere in the middle ($P \succ Q \succ R$). If your preferences satisfy the rule of continuity, then there exists some probability w such that $Q \sim wP + (1-w)R$. In other words, there must be some probability w where you indifferent between a w-weighted coin flip for P and R, and the "split the difference" option Q.

(4) Your preference for P or Q remains unchanged in the face of a third option. Formally, if $P \succ Q$, w is a probability, and R is any third lottery, then

$$wP + (1-w)R \succ wQ + (1-w)R\,.$$

That is, you prefer P to Q, then you also prefer a lottery involving P to the same lottery involving Q.

The significance of these rules is that they make possible the construction of a *utility function*, which has the following significance. Suppose that an agent's preferences among outcomes satisfy Rules 1–4. Then there exists a utility function $u(A_i)$ that assigns a real number to each possible outcome A_i, and that satisfies two properties:

1. For any two lotteries P and Q and any probability w,

$$u\{wP + (1-w)Q\} = w\,u(P) + (1-w)u(Q)$$

2. For any two lotteries P and Q, $E\{u(P)\} > E\{u(Q)\}$ if and only if $P \succ Q$.

By $E\{u(P)\}$ we mean the expected utility of lottery P, recalling that P is a random variable described by some probability distribution.

To sum it up:

1. The principle of expected utility says that, when presented with a choice among options whose outcomes are uncertain, take the option with the highest expected utility.
2. Thus to compare options, we need to be able to compute expected utilities.
3. To compute expected utilities, we need to be able to compute the expected value of a function (specifically, your utility function) of a random variable.

Usually step 3 is the hard part. It often necessitates a Monte Carlo simulation.

7 Monte Carlo

Basic idea. The idea of Monte Carlo simulation is to approximate complicated probability distributions via computer simulations. The key equation is Monte Carlo simulation is the following. If X is a random variable with probability distribution $P(X)$, and we are interested in computing the expected value of some function $f(X)$, then

$$E[f(X)] \approx \frac{1}{N}\sum_{i=1}^{N} f(X^{(i)}),$$

where each $X^{(i)}$ is a simulated draw from the distribution $P(X)$. This statement essentially says that we can approximate a population mean (the expected value on the left) with a sample mean

(the average on the right). The number N is called the Monte Carlo sample size. The Monte Carlo error is the discrepancy between the left-hand side and the approximation on the right. Larger values of N will lead to smaller Monte Carlo error.

Joint distributions. The Monte Carlo method works for joint distributions, too. Suppose that $X_1,\ldots,X_D$ are D correlated random variables with joint distribution $P(X_1,\ldots,X_D)$. For example, the random variables might be the returns on a single day of all the assets in a financial portfolio. If we're interested in some nonlinear function $f(X_1,X_2,\ldots,X_D)$ of these random variables, then we can use the same basic equation as above:

$$E[f(X_1,X_2,\ldots,X_D)] \approx \frac{1}{N}\sum_{i=1}^{N} f(X_1^{(i)},X_2^{(i)},\ldots,X_D^{(i)}),$$

where each set $(X_1^{(i)},X_2^{(i)},\ldots,X_D^{(i)})$ is a single draw from the joint distribution.

Sampling complicated joint distributions. We've seen how it is often necessary to simulate from a complicated joint distribution $P(X_1,\ldots,X_D)$. This is sometimes difficult to do, because the joint distribution is very complicated mathematically, and we lack the ability to describe it in all its glory. In this situation, a very practical technique is bootstrap resampling. Suppose we have M past samples of the random variables of interest, stacked in a matrix or spreadsheet:

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1D} \\ X_{21} & X_{22} & \cdots & X_{2D} \\ \vdots & & & \\ X_{M1} & X_{M2} & \cdots & X_{MD} \end{pmatrix}$$

where X_{ij} is the ith sample of the jth variable. For example, the ith row of this spreadsheet might give the returns/interest rates of D correlated assets on a single day.

The key idea of bootstrap resampling is the following. We may not be able to describe what the joint distribution $P(X_1,\ldots,X_D)$ is, but *we do know that every row of this X matrix is a sample from this joint distribution.* Therefore, instead of sampling from the joint distribution, we will sample from the sample—i.e. we will bootstrap the past data. Every time we need a new draw from the joint distribution $P(X_1,\ldots,X_D$, we randomly sample (with replacement) a single row of X.