Abstract

Two aspects of Predictive Analysis will be considered: choosing a regression equation for prediction, and predictive distributions for future observations. Articles from 1920 and the 1960s will be discussed as precursors to modern “Predictive Analytics” for such problems.

In 1960, Nicholson expounded upon the fact of “shrinkage” in regression, as happens when a predicting function developed on training data is applied to future observations.
Also in 1960, Stein considered this problem and developed a criterion for choosing a regression equation for prediction, developing a criterion with an adjustment reminiscent of that in adjusted R-square.

Consider the problem of modeling a dataset of employee days ill, or accidents in a population of insureds. One can consider a spectrum of granularity in describing the population, from a single distribution, to a bimodal distribution, to a mixture of two or more distributions, where there are two or more subpopulations, to modeling the population at the individual level, where each individual in a way constitutes a subpopulation. Cluster Analysis, the Mixture Model, and Bayesian models will be considered as one moves from one end of this spectrum to the other. It is noted that, in 1920, Greenwood and Yule developed a model with a distribution across the population parameter which would now be called a Bayesian model. Such developments are reviewed in relation to the “new” Predictive Analytics.

1 Introduction

It is planned in the first part of the course, among other things, to discuss two particular examples of Bayesian modeling, the Gamma-Poisson model and the Beta-Binomial model. “Gamma-Poisson” means that the distribution of $X$ given $\lambda$ is Poisson, and the prior distribution on the parameter $\lambda$ is Gamma. “Beta-Binomial” means that the distribution of $X$ given the success probability $\pi$ is Binomial, and the prior distribution on the parameter $\pi$ is Beta. In later notes, we’ll take up the Beta-Binomial model.

2 Various Models

We’ll begin by viewing the dataset of days ills for a sample of 50 miners. The number ranges for 0 to 18, with various frequencies. First, histograms are viewed. (See the accompanying Excel dataset in this folder.)

| Days ill | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|----------|---|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|
| Frequency| 2 | 3 | 5 | 5 | 2 | 5 | 5 | 4 | 6 | 3 | 0  | 1  | 4  | 1  | 2  | 0  | 0  | 1  | 1  |

We’ll fit a single Poisson distribution, and then a mixture of two. This begs the question of whether maybe three might be needed. Then, extrapolating this reasoning, we could conceive of modeling the situation in terms of each individual’s having his or her own value of the Poisson parameter. This is the topic of this lecture.

2.1 Fitting Two Distributions to the days ill dataset

min = 0, max = 18, mean = 6.6 days ill
2.1.1 minimum WGSS
Optimal means by this method: 4.9, 13.5

2.1.2 Maximum square of unequal-variance Student’s \(t\)
Optimal means by this method: 2.1, 8.9 days ill

2.1.3 K-means clustering
Should give the same result as min WGSS.

2.2 Finite Mixture Model
Suppose the population \(\Pi\) consists of two sub-populations, \(\Pi_1, \Pi_2\) Then the p.m.f. for \(X\) is
\[
p(x) = \pi_1 p_1(x) + \pi_2 p_2(x).
\]
To see this, note that
\[
\Pr\{X = x\} = \Pr(\text{individual is from } \Pi_1 \text{ and gives the value } x) + \Pr(\text{individual is from } \Pi_2 \text{ and gives the value } x)
= \Pr(\Pi_1) \Pr(x | \Pi_1) + \Pr(\Pi_2) \Pr(x | \Pi_2)
= \pi_1 p_1(x) + \pi_2 p_2(x).
\]

In the second lecture or so, we’ll look at the corresponding likelihood for a sample of \(N\) and the resulting equations for the Maximum Likelihood Estimators.

2.3 A Bayesian Model: An Example of Averaging over a Population
Putting a population distribution over a parameter can be a very helpful way of modeling. Here is an example. It is financial, in fact, actuarial. It is a model for accident rates in a population. Suppose that the yearly number of accidents of any given individual \(i\) in a population is distributed according to a Poisson distribution with parameter \(\lambda_i\) accidents per year. (Note that \(\lambda\) may be rather small. That would be nice!)

Then the probability that an individual with parameter value \(\lambda\) has exactly \(k\) accidents in a year, \(k = 0, 1, 2, \ldots\), is
\[
e^{-\lambda} \frac{\lambda^k}{k!}.
\]
Some individuals are more accident prone (have a higher accident rate) than others, so different individuals have different values of \(\lambda\).

2.3.1 Prior distribution on \(\lambda\)
A distribution can be put on \(\lambda\) to deal with this.
Suppose that this distribution is related to the gamma distribution with parameter \( m > 0 \), with p.d.f.

\[
f(\lambda) = \text{Const. } \lambda^{m-1} e^{-\lambda}, \lambda > 0.
\]

The constant is \( 1/\Gamma(m) \), where

\[
\Gamma(m) = \int_0^\infty \lambda^{m-1} e^{-\lambda} \, d\lambda.
\]

For \( m \) a positive integer, \( \Gamma(m) = (m - 1)! \). More generally, we can consider the two-parameter Gamma family with shape parameter \( m \) and scale parameter \( \beta > 0 \):

\[
f(\lambda; m, \beta) = \lambda^{m-1} e^{-\lambda/\beta} / \Gamma(m) \beta^m, \lambda > 0.
\]

Remark. For \( m = n/2 \) and \( \beta = 2 \), this is the chi-square distribution with \( n \) degrees of freedom.

### 2.3.2 Marginal distribution of \( X \)

Then the marginal distribution of \( X \), the number of accidents that a randomly selected individual has in a year, is

\[
f_X(x) = \int_0^\infty f_{X,\lambda}(x, \lambda) \, d\lambda
\]

\[
= \int f_{X|\lambda} f_{\lambda}(\lambda) \, d\lambda
\]

\[
= \int_0^\infty e^{-\lambda} \frac{\lambda^x}{x!} \lambda^{m-1} e^{-\lambda/\beta} \, d\lambda / [\beta^m \Gamma(m)].
\]

Working with this integral, this distribution can be shown to be a negative binomial. The negative binomial distribution with parameters \( m \) and \( p \) has p.m.f.

\[
p(x) = C(m + x - 1, m) p^m q^x, \quad x = 0, 1, 2, \ldots, \text{ where } q = 1 - p,
\]

where the symbol \( C(n, r) \) is the number of combinations of \( n \) things taken \( r \) at a time; \( C(n, r) = n! / r!(n - r)! \). This distribution is the distribution of the number of Bernoulli trials above \( m \) required to obtain \( m \) successes. That is, \( X \) is the excess number of trials. The mean is \( mq/p \) and the variance is \( mq/p^2 \). In the present application, \( p = 1/(1 + \beta) \). The mean is \( m\beta \) and the variance is \( m\beta (1 + \beta) \).

What if we ask, “What is the distribution on the accident-proneness parameters? ” Well, the hyperparameters \( m, \beta \) of the prior Gamma distribution can be estimated, e.g., by the method of moments applied to a sample of \( N \) observations \( X_1, X_2, \ldots, X_N \) of \( X \). To do this, we set \( \bar{X} \) equal to \( mq/p \) and \( s^2 \) equal to \( mq/p^2 \) and solve for \( m \) and \( p \).

### 2.3.3 Posterior distribution of \( \lambda \)

Suppose an individual has a value \( x \), that is, has \( x \) accidents. What would be the estimate of the accident-proneness \( \lambda \) for this individual?

To answer this, we look at the posterior distribution of \( \lambda \), with p.d.f. denoted by \( f_{\lambda|X}(\lambda|x) \). It turns out to be a Gamma distribution, that is, it is in the same family as the prior. When the prior and the posterior are in the same family, that family is said to be a conjugate distribution for the distribution of \( X \) given the parameter (Poisson in this case).
2.4 \emph{t} Distribution as a Scale-Mixture of Normals

The “Student’s” \emph{t} distribution can be obtained as a mixture of Normal distributions. This mixture is one of distributions with different variances. Such a mixture is called a \emph{scale} mixture. To develop this, let \( T \) and \( U \) be random variables, \( T \) given \( U = u \) being distributed according to according to a Normal distribution and \( U \) according to a chi-square distribution with \( m \) d.f. The conditional distribution of \( T \) given \( U = u \) is taken as Normal with mean zero and variance \( m/u \); that is, \( u/m \) is the reciprocal variance. If \( u \) is large, the conditional variance of \( T \) is small. The p.d.f. of \( U \) is

\[
f_U(u) = cu^{m/2-1}e^{-u/2}, \quad u > 0,
\]

where the constant \( c = \Gamma(m/2) / 2^{m/2} \). One can then write the conditional p.d.f. of \( T \), given that \( U = u \), derive the (unconditional) p.d.f. of \( T \) as

\[
f_T(t) = \int_0^\infty f_{T,U}(t,u) \, du = \int_0^\infty f_{T|U}(t|u) f_U(u) \, du,
\]

and verify that it is a Student’s \( t \) distribution, the p.d.f. of which is

\[
f_T(t) = \text{Const.} \, (1 + t^2/m)^{-(m+1)/2},
\]

where \( \text{Const.} = \Gamma(m+1) / \sqrt{m\pi} \Gamma(m/2) \).

This way of deriving the \( t \) distribution is summarized by saying that \( t \) distributions are \emph{scale-mixtures} of Normal distributions.

The members of the family of \( t \) distributions have heavier tails than Normal distributions and so can be useful for modeling various variables, such as RORs.

3 Exercises

3.1 General Exercises

1. Assume that, over the population, \( \lambda \) has an exponential distribution with mean \( 1/2 \).

\[
f(\lambda) = (1/2) \exp(-\lambda/2), \quad 0 < \lambda < \infty.
\]

Given that the conditional distribution given \( \lambda \) of the number of accidents \( X \) is Poisson with parameter \( \lambda \), what is the marginal distribution of the number \( X \) of accidents, giving the probability that a randomly selected individual has exactly \( k \) accidents in a year, \( k = 0, 1, 2, \ldots \)?

2. (continuation) What is the name of this distribution, and what is the parameter and its value in this case?

3. (continuation) Now suppose that \( \lambda \) has a Gamma distribution with shape parameter \( m \) and scale parameter \( \gamma \),

\[
f(\lambda; m, \gamma) = \text{Const.} \lambda^{m-1} \exp(-\lambda/\gamma).
\]

(The preceding case was \( m = 1, \gamma = 2 \).) Use the fact that the integral \( \int_0^\infty t^{m-1} e^{-t} \, dt \) is \( \Gamma(m) \), the gamma function with argument \( m \), to evaluate the constant.

4. (continuation) What is the marginal distribution, giving the probability that a randomly selected individual has exactly \( k \) accidents in a year, \( k = 0, 1, 2, \ldots \)?
5. What is the name of this distribution, and what is the parameter and its value in this case?

6. What is the posterior distribution of $\lambda$?

7. Estimate $\lambda$ for an individual who had $k = 2$ accidents in the year.

### 3.2 Theoretical Exercises

8. Carry out the indicated details of the derivation of the $t$ distribution as a scale mixture of Normal distributions.

9. If $U$ has a chi-square distribution with $m$ d.f., what is the p.d.f. of its reciprocal, $1/U$?

10. Show that as the number $m$ of degrees of freedom tends to infinity, the p.d.f. of the $t$ distribution converges pointwise to that of the standard Normal distribution.

11. Show that $\ln x \approx x - 1$ if $x$ is close to 1. Hint: The first terms of a Taylor series expansion of $f(x)$ about $x = a$ imply that $f(x) \approx f(a) + (x - a)f'(a)$ when $x$ is close to $a$. Take $f(x) = \ln x$ and $a = 1$.

### 4 Bibliography


Baranchik, A. J.


Donoho, David. Donoho. 50 Years of Data Science. Course Notes, Sept. 18, 2015. [courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf](courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf)


Sclove, Stanley L.