Probability and Statistics with R

Assignment 2 Problem 1

Problem 1

Suppose X denote the number of goals scored by home team in premier league. We can assume X is a random variable. Then we have to build the probability distribution to model the probability of number of goals. Since X takes value in $\mathbb{N} = \{0, 1, 2, \cdots\}$, we can consider the geometric progression sequence as possible candidate model, i.e.,

$$S = \{a, ar, ar^2, ar^3, \cdots\}.$$

But we have to be careful and put proper conditions in place and modify S in such a way so that it becomes proper probability distributions.

1. Figure out the necessary conditions and define the probability distribution model using S.

Solution-

Assumptions:

- To define a probability distribution, firstly we must have to assume that $0 < a < 1$ and $0 < r < 1$, and

$$\sum_{i=0}^{\infty} ar^i = 1$$

So that we have:

$$\sum_{i=0}^{\infty} r^i = \frac{1}{a}$$

$$1 + r + r^2 + \cdots = \frac{1}{a}$$

$$\frac{1}{1-r} = \frac{1}{a}$$

$$\therefore a = 1 - r$$

Hence, our probability distribution is,

$$X \sim Geometric(r)$$

where,$0 < r < 1$ and r denotes the probability of scoring a goal.

and probability mass function is,

$$P(X = i) = (1 - r)r^i \quad , \qquad i = 0, 1, 2, \cdots$$

Our model is modified as

$$S = \{(1-r), (1-r)r, (1-r)r^2, (1-r)r^3, \cdots\}$$

2. Check if mean and variance exists for the probability model.

Solution-

To find the mean of the modified model that follows geometric distribution-

$$
\begin{aligned}
E[X] &= \sum_{i=0}^{\infty} iar^i = a\sum_{i=0}^{\infty} ir^i \\
&= a\left[\sum_{i=0}^{\infty}(i+1)r^i - \sum_{i=0}^{\infty} r^i\right] \\
&= a\left[\left(1+2r+3r^2+\cdots\right)\right) - \left(1+r+r^2+\cdots\right)\right] \\
&= a\left[\frac{1}{(1-r)^2} - \frac{1}{1-r}\right] \\
&= a\left[\frac{r}{(1-r)^2}\right] = \frac{r}{1-r} \\
\therefore E[X] &= \frac{r}{1-r}
\end{aligned}
$$

Now to find variance of the modified geometric model,

$$Var[X] = E[X^2] - (E[X])^2 = E[X(X-1)] + E[X] - (E[X])^2$$

Let's first calculate $E[X(X-1)]$,

$$
\begin{aligned}
E[X(X-1)] &= \sum_{i=0}^{\infty} i(i-1)ar^i \\
&= a\sum_{i=1}^{\infty} i(i-1)r^i \\
&= a(2r^2 + 6r^3 + 12r^4 + \cdots) \\
&= 2ar^2(1 + 3r + 6r^2 + \cdots) \\
&= 2ar^2\frac{1}{(1-r)^3} = \frac{2r^2}{(1-r)^2} \qquad \because (a = 1-r) \\
\therefore Var[X] &= \frac{2r^2}{(1-r)^2} + \frac{r}{1-r} - \frac{r^2}{(1-r)^2} \\
\implies Var[X] &= \frac{r}{(1-r)^2}
\end{aligned}
$$

3. Can you find the analytically expression of mean and variance.

Solution- From the above result, we found the analytically expression of mean and variance-

$$E[X] = \frac{r}{1-r} \quad , \qquad Var[X] = \frac{r}{(1-r)^2}$$

4. From historical data we found the following summary statistics

mean	median	variance	total number of matches
1.5	1	2.25	380

Using the summary statistics and your newly defined probability distribution model find the following:

a. What is the probability that home team will score at least one goal?

Solution- First we calculate values of a and r from the given historical data:

$$E[X] = \frac{r}{1-r} = 1.5 \implies r = 0.6$$

Probability of at least one goal,

```
r = 0.6
geomp_0 = dgeom(x = 0, prob = 1-r, log = FALSE)
geomp_at1 = 1 - geomp_0
geomp_at1
```

```
## [1] 0.6
```

$$P(X \geq 1) = 0.6$$

b. What is the probability that home team will score at least one goal but less than four goal?

Solution-

```
geomp_123 = dgeom(x = c(1,2,3), prob = 1-r, log = FALSE)
geomp_123 = sum(geomp_123)
geomp_123
```

```
## [1] 0.4704
```

$$P(1 \leq X < 4) = 0.4704$$

5. Suppose on another thought you want to model it with off-the shelf Poisson probability models. Under the assumption that underlying distribution is Poisson probability find the above probabilities, i.e.,

Solution- If we try to fit Poisson probability model, we have to take $\lambda = 1.5$ as it is given that mean $= 1.5$ and for Poisson probability distribution $mean = \lambda$

Probability mass function is,

$$P(Y = i) = e^{-\lambda}\lambda^i \frac{1}{i!} \quad , \qquad i = 0, 1, 2, \cdots$$

a. What is the probability that home team will score at least one goal?

Solution-

```
poisp_0 = dpois(x = 0, lambda = 1.5, log = FALSE)
poisp_atl_1 = 1-poisp_0
poisp_atl_1
```

```
## [1] 0.7768698
```

$$\therefore P(Y \geq 1) = 0.7768698$$

b. What is the probability that home team will score at least one goal but less than four goal?

Solution-

```
poisp_123 = dpois(x = c(1, 2, 3), lambda = 1.5, log = FALSE)
poisp_123 = sum(poisp_123)
poisp_123
```

```
## [1] 0.7112274
```

$$P(1 \leq Y < 4) = 0.7112274$$

6. Which probability model you would prefer over another?

Solution- The value of variance for Geometric probability model is coming out to be $Var[X] = 3.75$ and for Poisson probability model it is $Var[Y] = 1.5$, for both the models variance is not near the variance of the given historical data i.e. 2.25, hence both the models are not a good fit for this data. However if we have to choose one, we would prefer Poisson probability model as its variance is closer to the given variance as compared to the variance of Geometric probability model. Moreover, to count the number of goals Poisson distribution is better than the Geometric distribution as it is generally used to find the number of failures before the first success.

7. Write down the likelihood functions of your newly defined probability models and Poisson models. Clearly mention all the assumptions that you are making.

Solution-

Geometric probability model

If $X_1, X_2, \cdots, X_n$ is an independent and identically distributed random sample following Geometric distribution with probability mass function,

$$P(X = x) = (1 - r)r^x \quad , \quad x = 0, 1, 2, \cdots$$

Likelihood function,

$$L(r|x) = \Pi_{i=1}^{n}(1 - r)r^{x_i}$$

Poisson probability model

If $Y_1, Y_2, \cdots, Y_n$ is an independent and identically distributed random sample following Poisson distribution with probability mass function,

$$P(Y = i) = e^{-\lambda}\lambda^i \frac{1}{i!} \quad , \quad i = 0, 1, 2, \cdots$$

Likelihood function,

$$L(\lambda|y) = \Pi_{i=0}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$