Tutorial 7 (Wk8): Probability Distributions & Linear Regression

1. Binomial Distribution Model Example

In a computer science module, suppose you take a multiple-choice question test with 10 questions, and each question has 5 answer choices (a, b, c, d, e). What is the probability you get exactly 4 questions correct? Use the **Binomial Distribution Model formula**.

- N (trials) $= 10$
- $K =$ get exactly 4 questions correct

Use p (success) $= \frac{1}{5}$
Use p (failure) $= \frac{4}{5}$

2. Poisson Distribution Model Examples

Example (i)

A manufacturer produces light-bulbs that are packed into boxes of 100. If quality control studies indicate that 0.5% of the light-bulbs produced are defective, what percentage of the boxes will contain:

(a) no defective bulbs?

(b) 2 or more defective bulbs?

Let the probability of a light-bulb being defective be $p = 0.005$ (0.5%). Since each box contains 100 light-bulbs, we can model the number of defective bulbs in each box using the Poisson distribution, with the expected number of defective bulbs given by $\lambda = np = 100 \times 0.005 = 0.5$.

(a) **Probability of no defective bulbs:**
Using the Poisson distribution formula, the probability of observing k defective bulbs is:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

For $k = 0$ (no defective bulbs),

$$P(X = 0) = \frac{0.5^0 e^{-0.5}}{0!} = e^{-0.5} \approx 0.6065$$

Therefore, approximately 60.65% of the boxes will contain no defective bulbs.

(b) **Probability of 2 or more defective bulbs:**
To find the probability of having 2 or more defective bulbs, we first calculate the probabilities for 0 and 1 defective bulb and then subtract their sum from 1:

$$P(X \geq 2) = 1 - (P(X = 0) + P(X = 1))$$

For $k = 1$,

$$P(X = 1) = \frac{0.5^1 e^{-0.5}}{1!} = 0.5e^{-0.5} \approx 0.3033$$

So,

$$P(X \geq 2) = 1 - (0.6065 + 0.3033) = 1 - 0.9098 \approx 0.0902$$

Thus, approximately 9.02% of the boxes will contain 2 or more defective bulbs.

Example (ii)

Suppose it has been observed that, on average, 180 cars per hour pass a specified point on a particular road in the morning rush hour. Due to impending roadworks, it is estimated that congestion will occur closer to the city centre if more than 5 cars pass the point in any one minute. What is the probability of congestion occurring?
Let the average number of cars passing the point per minute be λ. Since it is observed that 180 cars pass per hour, the average rate per minute is:

$$\lambda = \frac{180}{60} = 3 \text{ cars per minute}$$

We can model the number of cars passing in a given minute using the Poisson distribution with $\lambda = 3$.

The question asks for the probability of more than 5 cars passing the point in one minute, which is $P(X > 5)$.

$$P(X > 5) = 1 - \sum_{k=0}^{5} P(X = k)$$

Using the Poisson probability formula:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Calculating $P(X = k)$ for $k = 0$ to 5:

$$
\begin{aligned}
P(X = 0) &= \frac{3^0 e^{-3}}{0!} = e^{-3} \approx 0.0498 \\
P(X = 1) &= \frac{3^1 e^{-3}}{1!} = 3 \cdot e^{-3} \approx 0.1494 \\
P(X = 2) &= \frac{3^2 e^{-3}}{2!} = 4.5 \cdot e^{-3} \approx 0.2240 \\
P(X = 3) &= \frac{3^3 e^{-3}}{3!} = 4.5 \cdot e^{-3} \approx 0.2240 \\
P(X = 4) &= \frac{3^4 e^{-3}}{4!} = 3.375 \cdot e^{-3} \approx 0.1680 \\
P(X = 5) &= \frac{3^5 e^{-3}}{5!} = 2.025 \cdot e^{-3} \approx 0.1008
\end{aligned}
$$

Summing these probabilities:

$$\sum_{k=0}^{5} P(X = k) \approx 0.0498 + 0.1494 + 0.2240 + 0.2240 + 0.1680 + 0.1008 = 0.9160$$

Therefore,

$$P(X > 5) = 1 - 0.9160 = 0.0840$$

Thus, the probability of congestion occurring is approximately 8.40%.

3. Multiple Linear Regression

Unlike simple linear regression, multiple linear regression allows more than two independent variables to be considered. The goal is to estimate a variable based on several other variables. The variable to be estimated is called the response or dependent variable. The variables that are used for the prediction are called explanatory, or independent variables (predictors).

Go through this excellent introduction to Multiple Linear Regression:

https://www.investopedia.com/terms/m/mlr.asp

Multiple Linear Regression

and then try to solve the following example:
Suppose the following houses are for sale:

- A 2-bedroom house with 1 bathroom costs £150,000
- A 3-bedroom house with 1 bathroom costs £200,000

- A 2-bedroom house with 2 bathrooms costs £180,000

Build a multilinear prediction model for the house price based on its number of bedrooms and bathrooms.

Let the house price be predicted based on the number of bedrooms (x_1) and the number of bathrooms (x_2) using a linear regression model:

$$\text{Price} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where: - β_0 is the intercept, - β_1 is the coefficient for the number of bedrooms, and - β_2 is the coefficient for the number of bathrooms.

Given data:

- A 2-bedroom, 1-bathroom house costs £150,000.
- A 3-bedroom, 1-bathroom house costs £200,000.
- A 2-bedroom, 2-bathroom house costs £180,000.

We can set up the following system of equations:

$$\begin{cases} \beta_0 + 2\beta_1 + 1\beta_2 = 150000 \\ \beta_0 + 3\beta_1 + 1\beta_2 = 200000 \\ \beta_0 + 2\beta_1 + 2\beta_2 = 180000 \end{cases}$$

Subtracting the first equation from the second:

$$(\beta_0 + 3\beta_1 + \beta_2) - (\beta_0 + 2\beta_1 + \beta_2) = 200000 - 150000$$

$$\beta_1 = 50000$$

Now substitute $\beta_1 = 50000$ into the first equation:

$$\beta_0 + 2(50000) + \beta_2 = 150000$$

$$\beta_0 + 100000 + \beta_2 = 150000$$

$$\beta_0 + \beta_2 = 50000$$

Now use the third equation:

$$\beta_0 + 2(50000) + 2\beta_2 = 180000$$

$$\beta_0 + 100000 + 2\beta_2 = 180000$$

$$\beta_0 + 2\beta_2 = 80000$$

Subtracting $\beta_0 + \beta_2 = 50000$ from $\beta_0 + 2\beta_2 = 80000$:

$$\beta_2 = 30000$$

Finally, substitute $\beta_2 = 30000$ into $\beta_0 + \beta_2 = 50000$:

$$\beta_0 = 50000 - 30000 = 20000$$

Thus, the model is:

$$\text{Price} = 20000 + 50000 \cdot x_1 + 30000 \cdot x_2$$