Principles of Statistical Data Analysis

Home Work 1

Jiacheng Shen Sajjad Ayashm

2024-10-28

1 Data Exploration

In this section, we explore the data using summary statistics and visualizations to understand the distribution of ant abundance and soil moisture. Table 1 shows 5 rows of our data.

Table 1: Data in ants.RData file

abundance	moisture
2	20.0
226	17.6
52	12.2
37	15.6
0	20.5
255	20.5

1.1 Data Summary:

Table 2 shows the first few rows of the data set.

Table 2: Summary of Abundance of Ants and Moisture of Soil

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Abundance	0.00	6.75	35.50	86.31	99.25	767.00
Moisture	12.0	15.4	17.9	17.5	19.9	21.5

Additionally, the variance of ant abundance in the sample is **1.99e+04**, and for moisture, it is **7.083**

1.2 Data Visualization

We visualized the distribution of ant abundance using a histogram (Figure 1) and boxplot (Figure 2) to identify its spread and any potential outliers. A scatter plot of soil moisture versus ant abundance (Figure 3) was created to explore their relationship. Additionally, a mean-variance plot (Figure 4) was used to assess overdispersion, which supports the choice of a negative binomial model for this data.

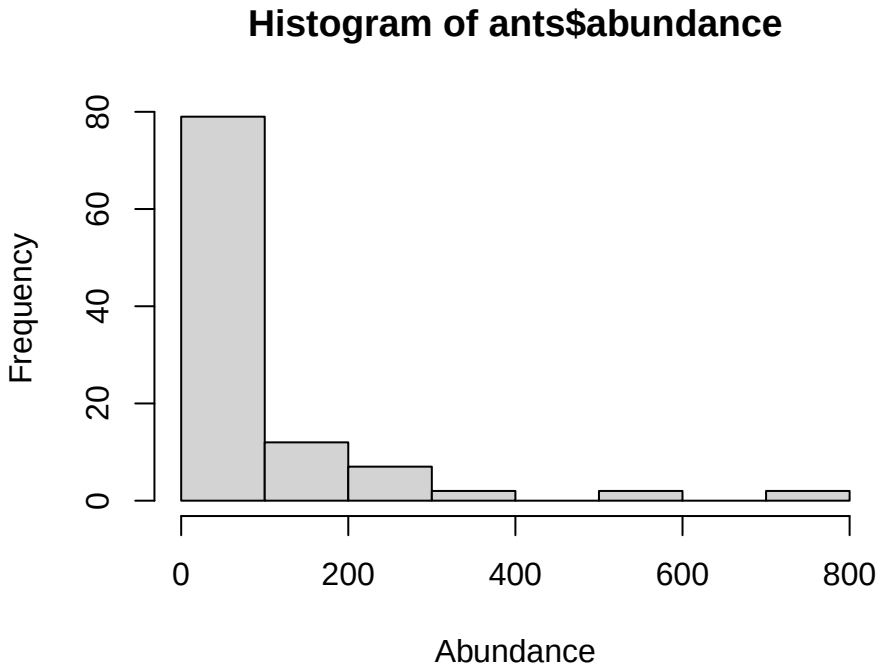

Figure 1: Histogram of Ant Abundance

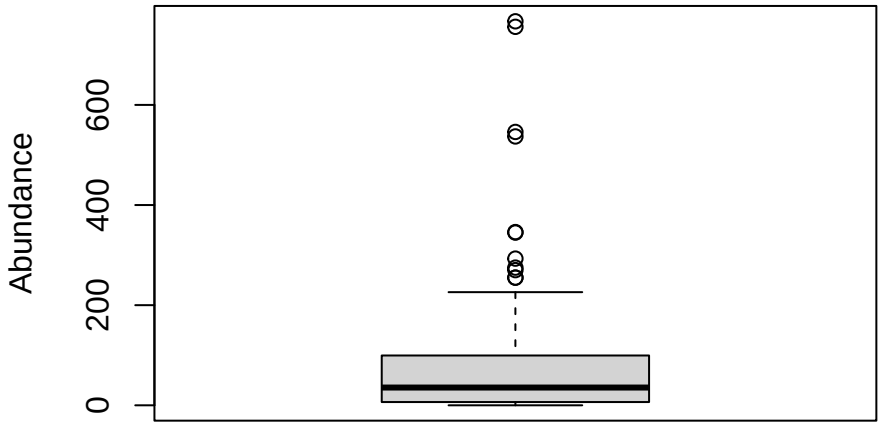

Figure 2: Boxplot of Abundance

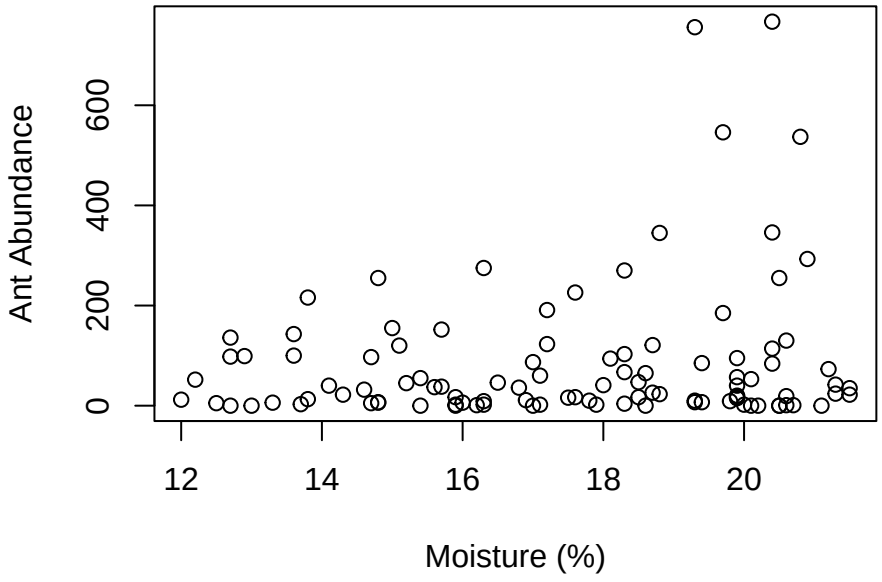

Figure 3: Scatter Plot of Moisture vs Ant Abundance

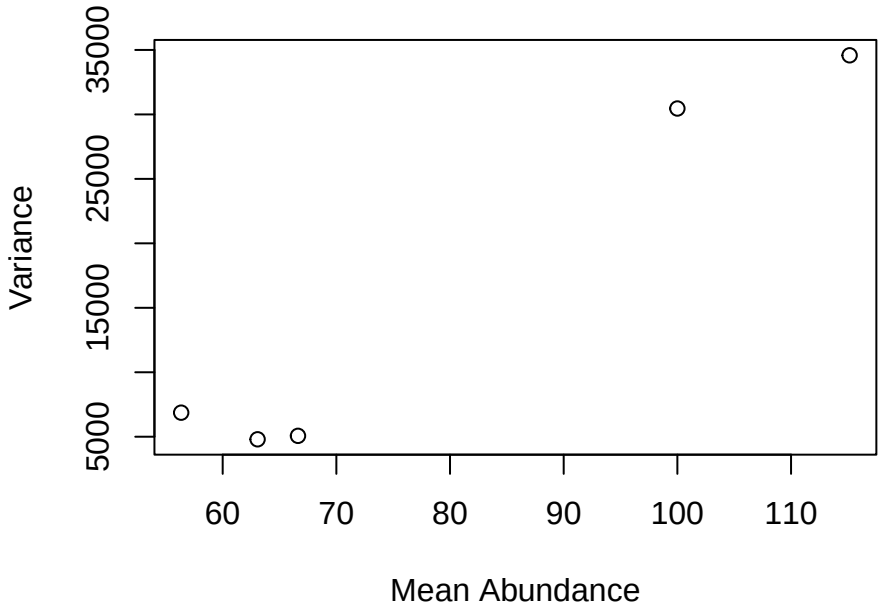

Figure 4: Mean vs Variance of Ant Abundance

2 Likelihood Function

The likelihood function is shown in Equation.

$$L(\beta_0, \beta_1, \phi) = \prod_{i=1}^{n} \frac{\Gamma(y_i + 1/\phi)}{\Gamma(1/\phi)\, y_i!} \left(\frac{1}{1+\mu_i\phi}\right)^{1/\phi} \left(\frac{\mu_i\phi}{1+\mu_i\phi}\right)^{y_i}$$

2.1 Likelihood Function

To define the likelihood function in R, we use a function that computes the product of probabilities for all observations.

Note: If we use the set of parameters described in the example above, the likelihood value results 0. One possible reason is that the number is too small to exhibit. That's why we derive the following log likelihood function.

2.2 Log-Likelihood Function

The log-likelihood function computes the sum of the log-probabilities instead of multiplying the probabilities directly, providing more numerical stability.

Using the same set of parameters, the log likelihood values result -1148.303, which makes more sense.

Explanation: Likelihood vs. Log-Likelihood: The likelihood function directly computes the product of individual probabilities, which can lead to underflow when dealing with very small values. The log-likelihood function sums the logarithms of the probabilities, providing greater numerical stability.

3 Estimating Equation for β_1

Our objective is to derive the estimating equation for β_1 by maximizing the log-likelihood function. The maximum likelihood estimate (MLE) for β_1 can be obtained by setting the derivative of the log-likelihood with respect to β_1 to zero, which requires numerical optimization.

The log-likelihood function is given by:

$$\ell(\beta_0, \beta_1, \phi) = \sum_{i=1}^{n} \left[\log\left(\Gamma(y_i + \frac{1}{\phi})\right) - \log\left(\Gamma(\frac{1}{\phi})\right) - \log(y_i!) + \frac{1}{\phi}\log\left(\frac{1}{1+\exp(\beta_0+\beta_1 x_i)\phi}\right) + y_i \log\left(\frac{\exp(\beta_0+\beta_1 x_i)\phi}{1+\exp(\beta_0+\beta_1 x_i)\phi}\right) \right]$$

To find the estimating equation for β_1, we differentiate this log-likelihood with respect to β_1:

$$\ell(\beta_1) = \sum_{i=1}^{n} \left[\frac{1}{\phi}\log\left(\frac{1}{1+\exp(\beta_0+\beta_1 x_i)\phi}\right) + y_i \log\left(\frac{\exp(\beta_0+\beta_1 x_i)\phi}{1+\exp(\beta_0+\beta_1 x_i)\phi}\right) \right]$$

Differentiating both terms with respect to β_1 yields:

$$\frac{\partial \ell(\beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} \left[\frac{x_i\,(y_i - \exp(\beta_0+\beta_1 x_i))}{1+\exp(\beta_0+\beta_1 x_i)\phi} \right]$$

Setting this result equal to zero gives the estimating equation:

$$\sum_{i=1}^{n} \left[\frac{x_i \left(y_i - \exp(\beta_0 + \beta_1 x_i) \right)}{1 + \exp(\beta_0 + \beta_1 x_i)\phi} \right] = 0$$

3.1 Step 1: Negative Log-Likelihood Function

Since the `optim()` function in R minimizes functions, we use the negative log-likelihood function for the optimization process. The negative log-likelihood is simply the negative of the log-likelihood function.

3.2 Step 2: Perform Optimization to Estimate Parameters

We now use the `optim()` function to find the maximum likelihood estimates (MLE) for β_0, β_1, and ϕ. The initial guesses for the parameters are provided, and the optimization is performed using the BFGS method.

The estimated parameters provide insights into the relationship between soil moisture and ant abundance. The intercept (β_0) is 2.504, representing baseline abundance, while the coefficient for moisture (β_1) at 0.109 shows the effect of soil moisture. The overdispersion parameter (ϕ) at 2.288 captures additional variability in the data.

4 Log-Likelihood Optimization with Respect to β_1

In this question, we explore the relationship between β_1 and the log-likelihood function, given the parameter estimates for β_0 and ϕ. Our objective is to visualize how changes in β_1 affect the log-likelihood, helping us understand its influence on the model.

In the Figure 5 above, we observe the log-likelihood values across a range of β_1 values. The peak of this curve indicates the value of that maximizes the log-likelihood, corresponding to the most likely estimate ofβ_1 given the data. This visualization aids in understanding how sensitive the model's fit is to changes in β_1.

5 Estimating the Optimal Value of β_1

To estimate the optimal value of β_1 that maximizes the log-likelihood, we use the `optim()` function in R with the BFGS optimization method:

The estimated value of β_1 is 0.109, which provides the most likely parameter value based on the observed data and model assumptions.

6 Evaluating the Estimating Equation at β_1

To verify the estimated β_1 value from Question 5, we compute the estimating equation at β_1 to see if it approximates zero, as expected for a maximum likelihood estimate.

The resulting value of the estimating equation is **-0.017**. A result close to zero suggests that β_1 is indeed an optimal estimate, indicating that the observed data align well with the expected values under the model.

7 Visualizing the Mean-Variance Relationship

In this question, we use the estimated parameters to visualize the mean-variance relationship of predicted ant abundance and compare it to the observed data from Question 1.

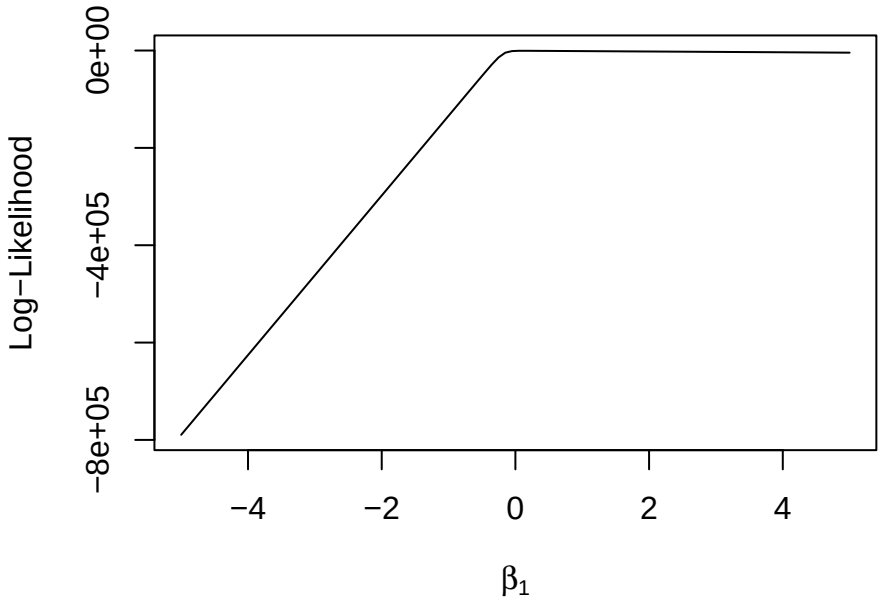

Figure 5: Log likelihood as a function of beta 1

7.1 Step 1: Calculating Predicted Mean and Variance

Using the model's estimated parameters, we calculate the predicted mean and variance of ant abundance for each soil moisture observation.

7.2 Step 2: Plotting the Predicted Mean-Variance Relationship

We plot the predicted mean against the predicted variance to visualize the model's mean-variance relationship for ant abundance.

7.3 Step 3: Comparing with Observed Data

To evaluate the model fit, we overlay the actual mean-variance relationship from the observed data (calculated in Question 1) on the same plot. (Figure 6)

7.4 Interpretation

The plot shows the relationship between the mean and variance of the predicted abundance (in blue) and the actual observed data (in red). If the predicted values closely match the observed data, this suggests that the model's mean-variance relationship aligns well with the data, supporting the negative binomial model's appropriateness.

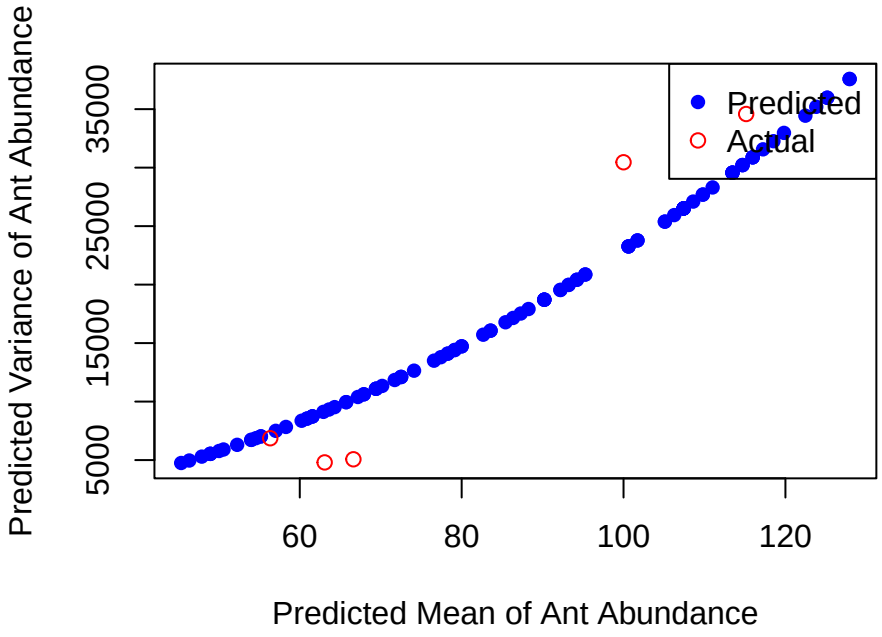

Figure 6: Mean-Variance Relationship of Predicted Ant Abundance