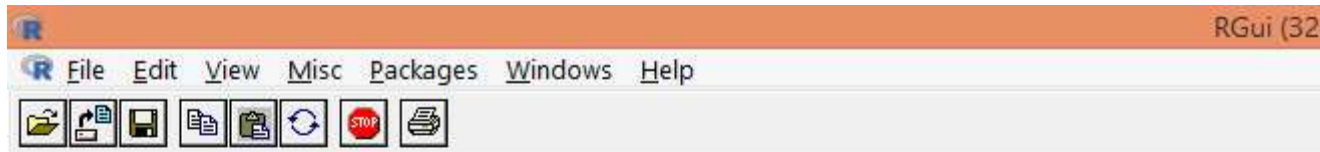


7. Having done with the preliminary analysis, we'll now apply Poisson regression as shown below



```

> fit <- glm(Species ~ Endemics + Area + Elevation + Nearest + Scrutz + Adjacent,
+           data = gale, family = poisson())
> summary(fit)

Call:
glm(formula = Species ~ Endemics + Area + Elevation + Nearest +
    Scrutz + Adjacent, family = poisson(), data = gale)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.9919  -2.9305  -0.4296   1.3254   7.4735

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.828e+00  5.958e-02  47.471  < 2e-16 ***
Endemics     3.388e-02  1.741e-03  19.459  < 2e-16 ***
Area        -1.067e-04  3.741e-05  -2.853  0.00433 **
Elevation    2.638e-04  1.934e-04   1.364  0.17264
Nearest      1.048e-02  1.611e-03   6.502  7.91e-11 ***
Scrutz       -6.835e-04  5.802e-04  -1.178  0.23877
Adjacent     4.539e-05  4.800e-05   0.946  0.34437
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

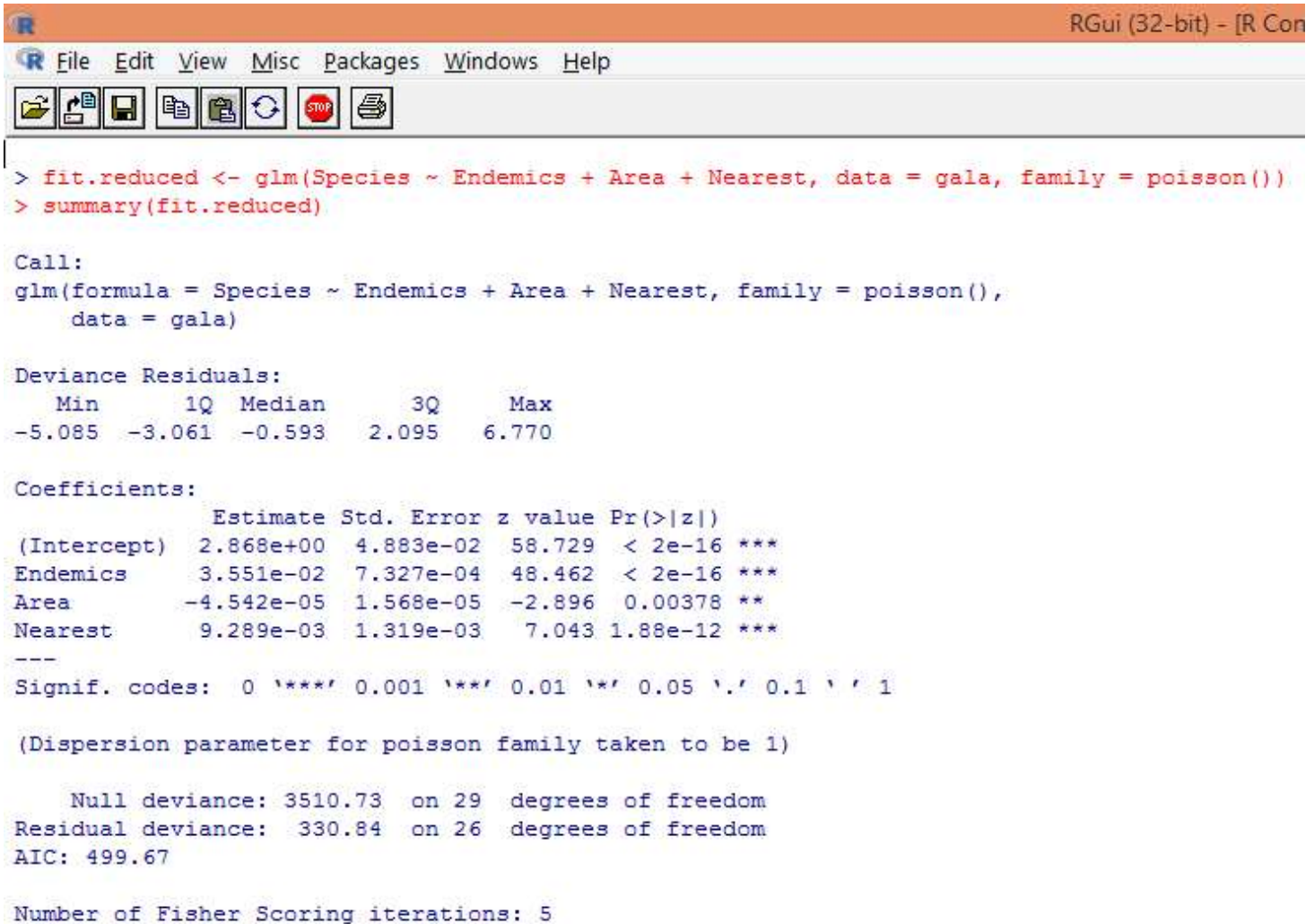
    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  313.36  on 23  degrees of freedom
AIC: 488.19

Number of Fisher Scoring iterations: 5

```

Based on the above analysis, we find that variables Endemics, Area, and Nearest are significant and only their inclusion is sufficient to build the right Poisson regression model.

8. We'll build a modified Poisson regression model taking into consideration three variables only viz. Endemics, Area, and Nearest. Let's see what results we get.



```

RGui (32-bit) - [R Con
File Edit View Misc Packages Windows Help
[Icons]

> fit.reduced <- glm(Species ~ Endemics + Area + Nearest, data = gala, family = poisson())
> summary(fit.reduced)

Call:
glm(formula = Species ~ Endemics + Area + Nearest, family = poisson(),
    data = gala)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.085  -3.061  -0.593   2.095   6.770

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.868e+00  4.883e-02  58.729 < 2e-16 ***
Endemics     3.551e-02  7.327e-04  48.462 < 2e-16 ***
Area        -4.542e-05  1.568e-05  -2.896  0.00378 **
Nearest      9.289e-03  1.319e-03   7.043 1.88e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

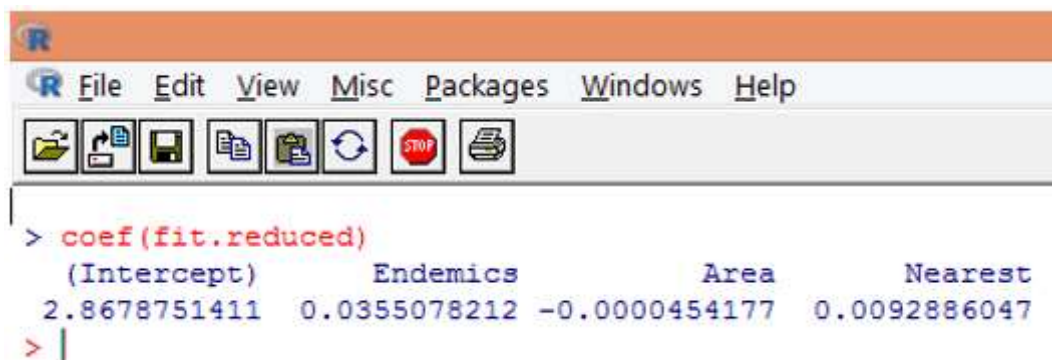
    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  330.84  on 26  degrees of freedom
AIC: 499.67

Number of Fisher Scoring iterations: 5

```

The output produces deviances, regression parameters, and standard errors. We can see that each of the parameters is significant at $p < 0.05$ level.

9. The next step is to interpret the model parameters. The model coefficients can be obtained either by examining Coefficients in the above output or by using `coef()` function.



```

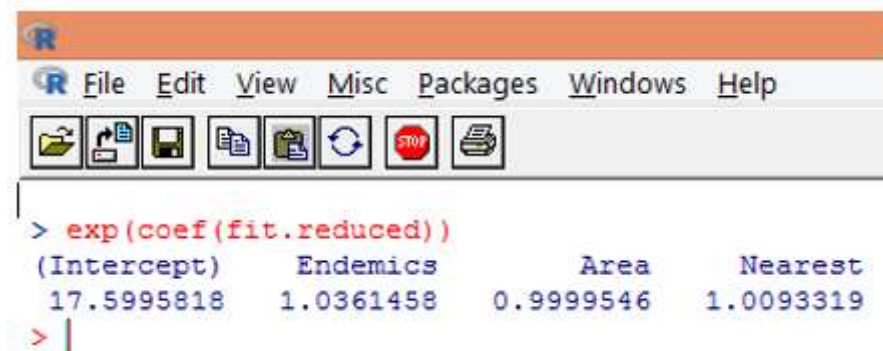
RStudio
File Edit View Misc Packages Windows Help
[Icons]

> coef(fit.reduced)
      (Intercept)      Endemics        Area        Nearest 
2.8678751411    0.0355078212 -0.0000454177  0.0092886047 
> |

```

In Poisson regression, the dependent variable is modeled as the log of the conditional mean $\log_e(l)$. The regression parameter of 0.0355 for Endemics indicates that a one-unit increase in the variable is associated with a 0.04 increase in the log mean number of Species, holding other variables constant. The intercept is a log mean number of Species when each of the predictors equals zero.

10. However, it is much easier to interpret the regression coefficients in the original scale of the dependent variable (number of Species, rather than log number of Species). The exponentiation of the coefficients will allow an easy interpretation. This is done as follows.

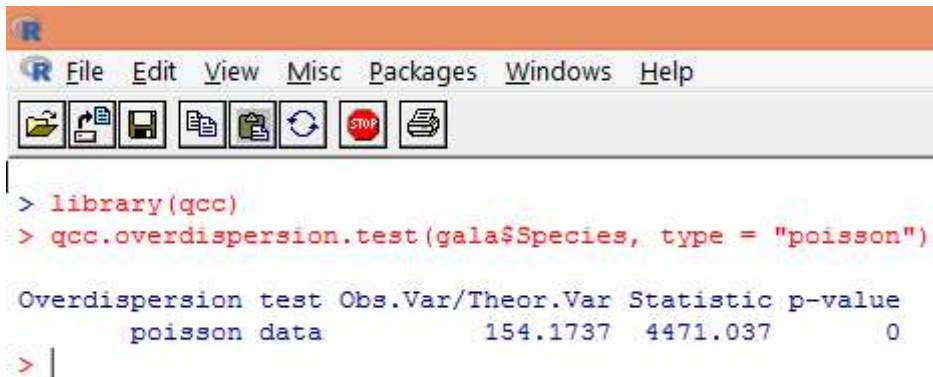


```
> exp(coef(fit.reduced))
(Intercept)    Endemics        Area    Nearest
 17.5995818    1.0361458    0.9999546    1.0093319
> |
```

From the above findings, we can say that one unit increase in Area multiplies the expected number of species by 0.9999, and a unit increase in the number of endemic species represented by Endemics multiplies the number of species by 1.0361. The most important aspect of Poisson regression is that exponentiated parameters have a multiplicative rather than an additive effect on the response variable.

11. Using the above steps, we obtained a Poisson regression model for predicting the number of plant species on the Galapagos Islands. However, it is very important to check for overdispersion. In Poisson regression, the variance and means are equal.

Overdispersion occurs when the observed variance of the response variable is larger than would be predicted by the Poisson distribution. Analyzing overdispersion becomes important as it is common with count data, and can negatively impact the final results. In R, overdispersion can be analyzed using the "qcc" package. The analysis is illustrated below.



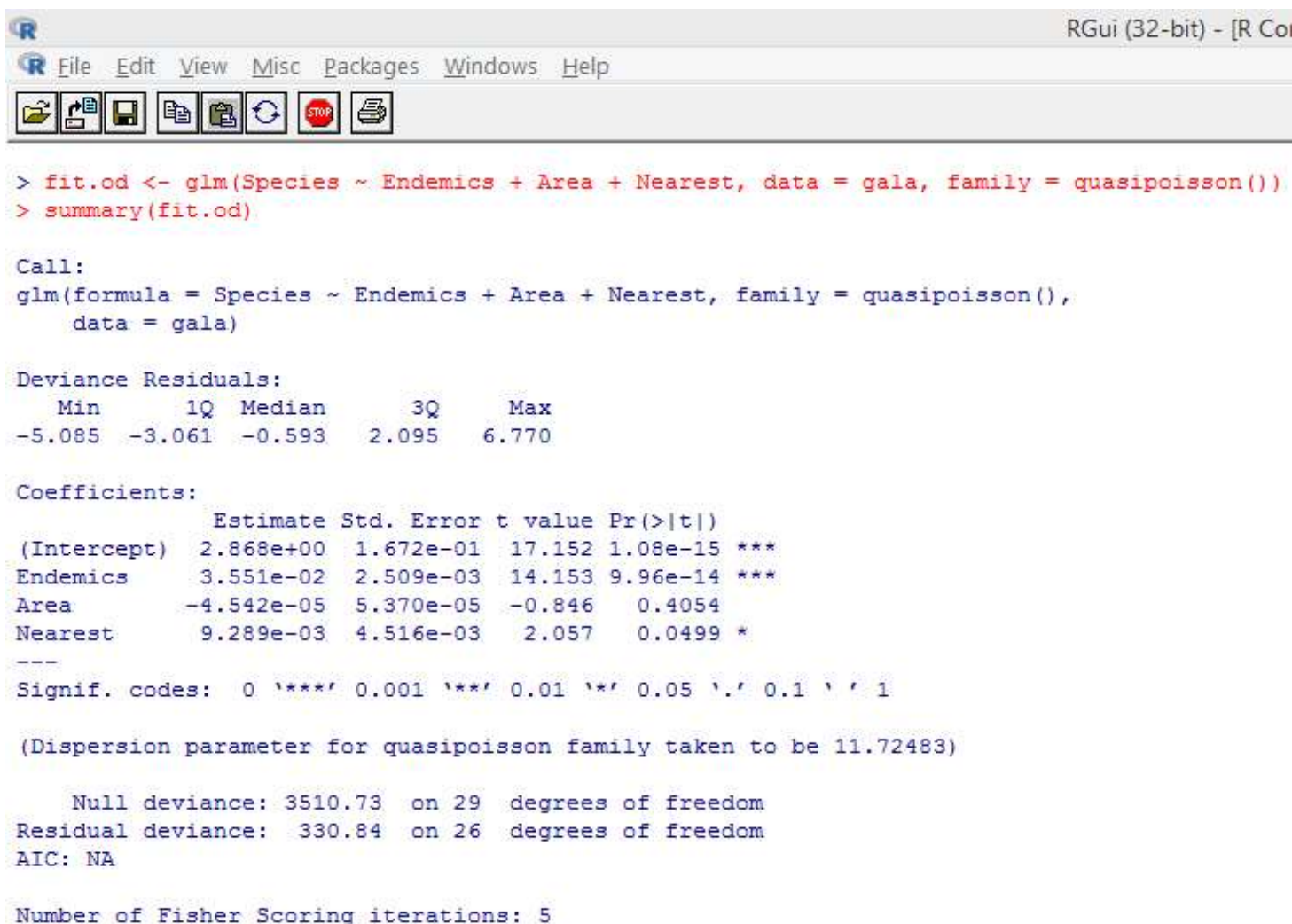
```

> library(qcc)
> qcc.overdispersion.test(gala$Species, type = "poisson")

Overdispersion test Obs.Var/Theor.Var Statistic p-value
poisson data          154.1737  4471.037      0
> |

```

The above significant test shows that the p-value is less than 0.05, which strongly suggests the presence of overdispersion. We'll try fitting a model using `glm()` function, by replacing family = "Poisson" with family = "quasipoisson". This is illustrated below.



```

> fit.od <- glm(Species ~ Endemics + Area + Nearest, data = gala, family = quasipoisson())
> summary(fit.od)

Call:
glm(formula = Species ~ Endemics + Area + Nearest, family = quasipoisson(),
    data = gala)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.085  -3.061  -0.593   2.095   6.770

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.868e+00  1.672e-01  17.152 1.08e-15 ***
Endemics     3.551e-02  2.509e-03  14.153 9.96e-14 ***
Area        -4.542e-05  5.370e-05  -0.846  0.4054
Nearest      9.289e-03  4.516e-03   2.057  0.0499 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 11.72483)

Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  330.84  on 26  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5

```

Closely studying the above output, we can see that the parameter estimates in the quasi-Poisson approach are identical to those produced by the Poisson approach, though the standard errors are different for both the approaches. Moreover, in this case, for Area, the p-value is greater than 0.05

which is due to larger standard error.

Importance of Poisson Regression

- Poisson Regression in R is useful for correct predictions of the discrete / count variable.
- It helps us identify those explanatory variables which have a statistically significant effect on the response variable.
- Poisson Regression in R is best suitable for events of “rare” nature as they tend to follow a Poisson distribution as against common events that usually follow a normal distribution.
- It is suitable for application in cases where the response variable is a small integer.
- It has wide applications, as a prediction of discrete variables is crucial in many situations. In medicine, it can be used to predict the impact of the drug on health. It is heavily used in survival analysis like the death of biological organisms, failure of mechanical systems, etc.

Conclusion

Poisson regression is based on the concept of Poisson distribution. It is another category belonging to the set of regression techniques that combines the properties of both Linear as well as Logistic regressions. However, unlike Logistic regression which generates only binary output, it is used to predict a discrete variable.