W271 Summer 2022 Lecture Video Question Solutions Week 5

Contents

Week 5 Discrete Response Model Part 5

5.2 Part 1, Poisson Probability Model: Outline

Q: How would you decide on rejection of the null hypothesis?

Solution: A hypothesis test in the Poisson model tests $H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \mu_0$. We form the test statistic $Z_0 = \frac{\hat{\mu}-\mu_0}{SE(\hat{\mu})}$ where $Z_0 \sim N(0,1)$. We can compare Z_0 to the critical value of the standard normal distribution based on the selected significance level α. **If $|Z_0| > Z_{critical}$ then we reject the null hypothesis and conclude that we have enough statistical evidence to conclude that $\mu \neq \mu_0$.**

Alternatively we can calculate the confidence interval $\hat{\mu} \pm Z_{1-\frac{\alpha}{2}} SE(\hat{mu})$ and check whether μ_0 is in the interval. If it is not in the confidence interval for $\hat{\mu}$, then we can reject the hull hypothesis. This is equivalent to calculating the test statistic and comparing to the critical value.

5.3 An Example

Q: Note that there were no vehicles remaining in the intersection for more than one stoplight cycle. Why is the above feature important for the application of the Poisson model to the problem?

Solution: The poisson distribution models the number of events occuring in a fixed time interval and assumes these events occur with a constant expected rate and also that the time intervals are independent of one another. If cars remain at the intersection for more than one stoplight, then the number of cars in each interval i.e. stoplight cycle are no longer independent of one another. **The count of cars in the current interval may be affected by the previous intervals if cars that arrived then stayed longer into subsequent intervals, violating the independence assumption.**

5.3 An Example

Q: Answer each question based on the estimated model in the code below.

```r
#data can be downloaded from here: https://www.chrisbilder.com/categorical/programs_and_data.html
dat <- read.csv("~/Documents/Berkeley W271/Week 5 Discrete Response Model Part 5/HorseshoeCrabs.csv")

p.model <- glm(Sat ~ Width, data = dat, family = poisson(link = "log"))

summary(p.model)
```

Solution:

```
##
## Call:
## glm(formula = Sat ~ Width, family = poisson(link = "log"), data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8526  -1.9884  -0.4933   1.0970   4.9221
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.30476    0.54224  -6.095  1.1e-09 ***
## Width        0.16405    0.01997   8.216  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 567.88  on 171  degrees of freedom
## AIC: 927.18
##
## Number of Fisher Scoring iterations: 6
```

What happens to the estimated mean number of satellites as the width increases? **The coefficient on width from the model output above is positive (0.16), meaning that the $log(\mu)$ or log of number of satellites around the crab increases as the width increases. Because log is a monotonic transformation, or the fact that $e^{0.16} > 1$, as width increases the number of satellites also increases.

Perform a Wald test for a B1. To perform, a wald test for B_1 we can directly use the model output below since a wald test just compares $|\frac{\hat{\beta}}{SE(\hat{\beta})}|$ to the critical value $Z_{1-\frac{\alpha}{2}}$ for a specified significance level. **Since the Z value for the width coefficient is $8.2 > 2$, we reject the null hypothesis that it is zero.**

Perform a LRT for an explanatory variable. Because we only have one explanatory variable in this model, peforming a likelihood ratio test on it will result in comparing the model fitted above to the null model. We can therefore use the null and residual deviances from the model fit to conduct this test since $NullDeviance = 2(loglik(SaturatedModel) - loglik(NullModel))$ and $ResidualDeviance = 2(loglik(SaturatedModel) - loglik(CurrentModel))$, which implies $2(loglik(CurrentModel) - loglik(NullModel)) = NullDeviance - ResidualDeviance$. This is distributed χ^2_1 since we have one coefficient in the model beyond the intercept.

Because the p-value from this test is so low i.e. less than the traditional cutoff of 0.05, we reject the null hypothesis and have stong evidence that this model is better than the null model with just the intercept according to the likelihood ratio test.

```
attributes(p.model)
```

```
## $names
##  [1] "coefficients"      "residuals"         "fitted.values"
##  [4] "effects"           "R"                 "rank"
##  [7] "qr"                "family"            "linear.predictors"
## [10] "deviance"          "aic"               "null.deviance"
## [13] "iter"              "weights"           "prior.weights"
## [16] "df.residual"       "df.null"           "y"
## [19] "converged"         "boundary"          "model"
## [22] "call"              "formula"           "terms"
## [25] "data"              "offset"            "control"
## [28] "method"            "contrasts"         "xlevels"
##
## $class
## [1] "glm" "lm"
```

```
test.statistic <- p.model$null.deviance - p.model$deviance
p.value <- pchisq(q = test.statistic, df = 1, lower.tail = F)
p.value
```

```
## [1] 7.82755e-16
```

We also repeat the likelihood ratio test using the anova function to show the results are the same. Note the p-values match.

```
model1 <- glm(Sat ~ 1, data = dat, family = poisson(link = "log"))
model2 <- glm(Sat ~ Width, data = dat, family = poisson(link = "log"))

anova(model1, model2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Sat ~ 1
## Model 2: Sat ~ Width
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       172     632.79
## 2       171     567.88  1   64.913 7.828e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Find a Wald confidence interval for μ. Are there any worries about interval limits being outside of the appropriate numerical range? Note the question as written does not really make sense to code because the value and interval for μ depends on width. We can write the theoretical interval though to study its properties:

$CI\ for\ \hat{\mu}(x) : exp(\hat{\beta}_0 + \hat{\beta}_1 x) \pm \sqrt{\frac{exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{n}}$

Now let's suppose that width is zero so that $x = 0$. Then:

$CI\ for\ \hat{\mu}(0) : exp(\hat{\beta}_0 \pm \sqrt{\frac{exp(\hat{\beta}_0)}{n}}$.

This can actually be negative depending on the values for $\hat{\beta}_0$ and n, which is theoretically impossible for the Poisson model of counts.

Hence there is some worries for the confidence Wald intervals near 0 that their range may extend beyond the theoretical range, but practically speaking that is ok. However, this is why there are other adjusted intervals like the Score CI for Poisson models.

This is also why when creating confidence intervals for predictions in GLM models, we usually find the intervals on the untransformed scale (log scale in the case of Poisson regression and logit scale in the case of logistic regression) and then transform the resulting values.

We can see this for the specific model here below using the predict function.

```
alpha <- 0.05
Z <- qnorm(1 - alpha/2)

## Wald interval for mu is below 0 for Width = 0
wald.conf <- predict(p.model, type = "response", newdata = data.frame(Width = 0), se.fit = T)
c(wald.conf$fit - Z * wald.conf$se.fit, wald.conf$fit + Z * wald.conf$se.fit)
```

```
##            1            1
## -0.002304313  0.075720557
```

```
## Wald interval for mu on log scale for Width = 0 is ok
log.wald.conf <- predict(p.model, type = "link", newdata = data.frame(Width = 0), se.fit = T)
c(exp(log.wald.conf$fit - Z * log.wald.conf$se.fit), exp(log.wald.conf$fit + Z * log.wald.conf$se.fit))
```

```
##          1          1
## 0.01268251 0.10624757
```