UNIVERSITY of WEST FLORIDA

Negative Binomial Regression

Statistics for Data Science II

Introduction

In the last lecture, we learned that the Poisson distribution is appropriate for count data.

However, the Poisson distribution assumes that the mean is equal to the variance.

The negative binomial distribution is an alternative that relaxes Poisson's assumption.

The negative binomial regression model is as follows:

$$\ln(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

Note that this is the same model as Poisson, however, we are assuming a different underlying distribution.

Check Assumptions

Example:

Recall the horseshoe crab example from the last lecture. Let's determine if Poisson was appropriate – we will compare the mean and the variance.

```
mean(data$satellites_num)
```

```
## [1] 2.919075
```

```
var(data$satellites_num)
```

```
## [1] 9.912018
```

Because the variance is larger than the mean, we know the data is overdispersed.

Check Assumptions

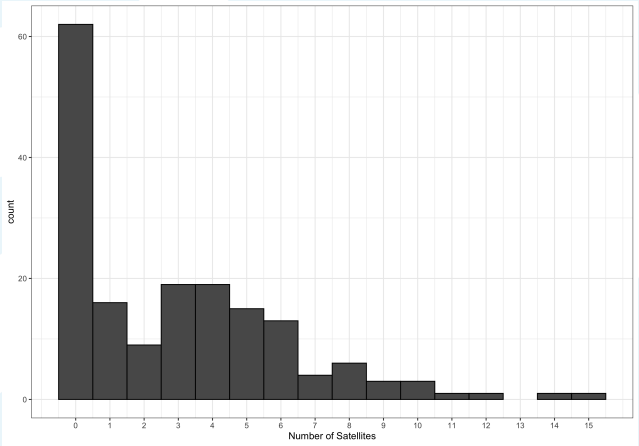
Example:

Let's also look at a histogram of the outcome to see if we can detect the overdispersion

```
p1 <- ggplot(data, aes(x=satellites_num)) +
  geom_bar(width = 1, color = "black") +
  scale_x_continuous(breaks=seq(0,15,1)) +
  xlab("Number of Satellites") +
  theme_bw()
```

Check Assumptions

Example:



Modeling

We will specify the negative binomial in R using the `glm.nb()` function.

e.g., `glm.nb(outcome ~ predictor1 + predictor2 + ..., data = dataset`

Example:

```
m1 <- glm.nb(satellites_num ~ width_cm + spine_cond +
             width_cm:spine_cond, data=data)
```

Modeling

Example:

```
summary(m1)[11]
```

```
## $coefficients
##                              Estimate Std. Error    z value  Pr(>|z|)
## (Intercept)            -0.76885871 3.25818739 -0.2359774 0.8134502
## width_cm                0.07630659 0.11992580  0.6362817 0.5245928
## spine_cond             -1.34216440 1.33021934 -1.0089798 0.3129843
## width_cm:spine_cond     0.04744452 0.04926015  0.9631421 0.3354762
```

The resulting model is

$$\ln(Y) = -0.77 + 0.08\text{width} - 1.34\text{spine} + 0.05(\text{width} \times \text{spine})$$

Comparing to Poisson

Example:

```
summary(poi)[12]
```

```
## $coefficients
##                          Estimate Std. Error    z value    Pr(>|z|)
## (Intercept)           0.26826395 1.46792355  0.1827506 0.854993689
## width_cm              0.03838017 0.05313290  0.7223429 0.470083674
## spine_cond           -1.58362166 0.61417266 -2.5784633 0.009924084
## width_cm:spine_cond   0.05614007 0.02235483  2.5113178 0.012028137
```

```
summary(nb)[11]
```

```
## $coefficients
##                          Estimate Std. Error    z value  Pr(>|z|)
## (Intercept)          -0.76885871 3.25818739 -0.2359774 0.8134502
## width_cm              0.07630659 0.11992580  0.6362817 0.5245928
## spine_cond           -1.34216440 1.33021934 -1.0089798 0.3129843
## width_cm:spine_cond   0.04744452 0.04926015  0.9631421 0.3354762
```

Interpretations, Inference, Diagnostics, etc.

Like under Poisson regression, we will convert the $\hat{\beta}$ to an IRR and interpret in terms of a multiplicative effect.

We construct the same Wald's z test for significant predictors using the `summary()` function.

We also construct the same confidence intervals using the `confint()` function.

We can obtain predicted values from the `predict()` function.

Finally, diagnostics are the same, too – we look at the plot for Cook's distance and check the VIF for multicollinearity.

Theoretical Considerations

The negative binomial regression includes a parameter which accounts for the overdispersion in the data.

This is how we relax the assumption from Poisson regression.

Why do we care about overdispersion?

When the data is overdispersed and we apply Poisson regression, we are underestimating the standard error.

When we underestimate the standard error, our test statistic ($\hat{\beta}$/stderr) becomes larger than it should be, which in turn causes the corresponding p-value to be smaller than it should be.

That is, we may determine there is a relationship more often than we should.

Theoretical Considerations

What if we use negative binomial when the mean is equal to the variance?

> As the dispersion parameter approaches 1 (mean = variance), the negative binomial converges in distribution to the Poisson.

Because it is easy to check the assumption, we can quickly make the determiniation between the two.

> However, it never "hurts" to use negative binomial regression over Poisson regression!