UNIVERSITY *of* WEST FLORIDA

Poisson Regression

Statistics for Data Science II

Introduction

Suppose we are faced with *count* data.

This is discrete data, not continuous.

Fortunately, the Poisson distribution is appropriate for count data.

The Poisson regression model is as follows:

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

Note that this is similar to logistic regression in that we are modeling the natural log of the outcome.

Modeling

We will specify this in R using the `glm()` function, specifying `family = "poisson"`.

e.g., `glm(outcome ∼ predictor1 + predictor2 + ..., data = dataset, family = "poisson")`

The Poisson distribution is used for count outcomes, thus, is why we specify it here.

Example:

```
m1 <- glm(satellites_num ~ width_cm + spine_cond +
            width_cm:spine_cond, family="poisson",
          data=data)
```

Modeling

Example:

```
summary(m1)[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
```

The resulting model is

$$\ln(Y) = 0.27 + 0.04\text{width} - 1.58\text{spine} + 0.06(\text{width} \times \text{spine}),$$

where y is the number of satellites a female horseshoe crab has

Interpretations

In Poisson regression, we convert the $\hat{\beta}_i$ values to incident rate ratios (IRR).

$$\text{IRR}_i = \exp\left\{\hat{\beta}_i\right\}$$

This is a multiplicative effect, like an odds ratio in logistic regression.

An IRR > 1 indicates an increase in the expected count.

An IRR < 1 indicates a decrease in the expected count.

We also interpret the IRR similar to the odds ratio:

For a 1 [unit of predictor] increase in [predictor name], the expected count of [outcome] is multiplied by [$e^{\hat{\beta}_i}$].

For a 1 [unit of predictor] increase in [predictor name], the expected count of [outcome] are [increased or decreased] by [100(e$^{\hat{\beta}_i}$-1)% or 100(1-e$^{\hat{\beta}_i}$)%].

Interpretations

Example:

Because our model contains an interaction, we must set one predictor (width or shell condition) to be constant. Then we can interpret the IRR for the other predictor. Let's look at a spine condition of 1 (best).

$$\begin{aligned} \ln(Y) &= 0.27 + 0.04\text{width} - 1.58\text{spine} + 0.06(\text{width} \times \text{spine}) \\ &= 0.27 + 0.04\text{width} - 1.58(1) + 0.06(\text{width} \times 1) \\ &= -1.31 + 0.10\text{width} \end{aligned}$$

Thus, the IRR $= \exp\{0.10\} = 1.11$.

When a female horseshoe crab has the best spine condition, for a 1 cm increase in shell width, we expect the number of satellites to increase by 11%.

UNIVERSITY of WEST FLORIDA

Interpretations

Example:

Because our model contains an interaction, we must set one predictor (width or shell condition) to be constant. Then we can interpret the IRR for the other predictor. Let's look at a shell width of 25 cm.

$$\begin{aligned} \ln(Y) &= 0.27 + 0.04\text{width} - 1.58\text{spine} + 0.06(\text{width} \times \text{spine}) \\ &= 0.27 + 0.0425 - 1.58\text{spine} + 0.06(25 \times \text{spine}) \\ &= 10.27 - 0.08\text{width} \end{aligned}$$

Thus, the IRR $= \exp\{-0.08\} = 0.92$.

When a female horseshoe crab has a shell width of 25 cm, for a 1 unit increase in shell condition (i.e., deteriorating spine), we expect the number of satellites to decrease by 8%.

Visualization of Interaction

Example:

```
min(data$width_cm)
```

```
## [1] 21
```

```
max(data$width_cm)
```

```
## [1] 33.5
```

```
shell_width <- seq(21, 34, 0.1)
head(shell_width)
```

```
## [1] 21.0 21.1 21.2 21.3 21.4 21.5
```

Visualization of Interaction

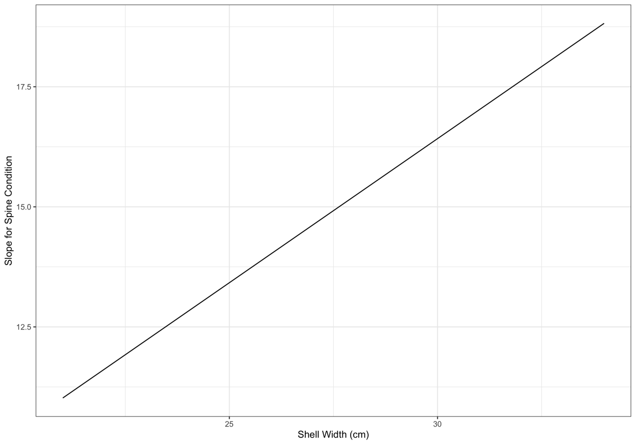
Example:

```
spine_slope <- -1.58 + 0.6*shell_width
graph <- tibble(shell_width, spine_slope)

p1 <- graph %>% ggplot(aes(x = shell_width, y = spine_slope)) +
  geom_line() +
  ylab("Slope for Spine Condition") +
  xlab("Shell Width (cm)") +
  theme_bw()
```

Visualization of Interaction

Example:



Visualization of Interaction

Example:

```
min(data$spine_cond)
```

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

```
max(data$spine_cond)
```

```
## [1] 3
```

```
spine <- seq(1, 3, 0.1)
head(spine)
```

```
## [1] 1.0 1.1 1.2 1.3 1.4 1.5
```

Visualization of Interaction

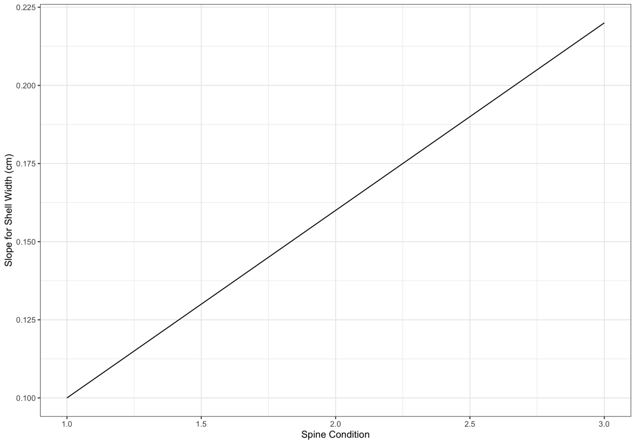
Example:

```r
shell_slope <- 0.04+0.06*spine
graph <- tibble(spine, shell_slope)

p2 <- graph %>% ggplot(aes(x = spine, y = shell_slope)) +
  geom_line() +
  ylab("Slope for Shell Width (cm)") +
  xlab("Spine Condition") +
  theme_bw()
```

Visualization of Interaction

Example:



Visualizations of Lines

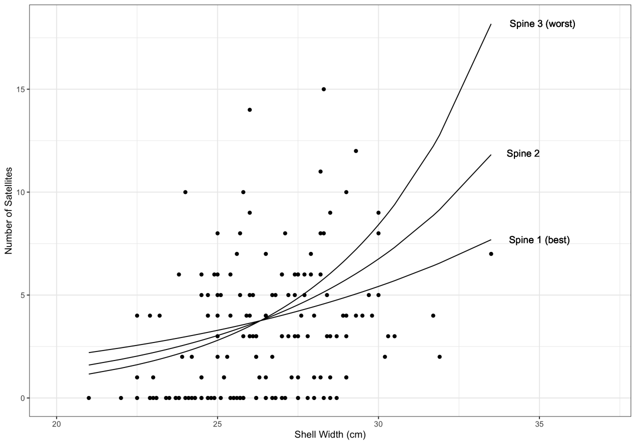
Example:

```
data <- data %>%
  mutate(exp_shell1 = exp(0.27+0.04*width_cm - 1.58*1 + 0.06*1*width_cm),
         exp_shell2 = exp(0.27+0.04*width_cm - 1.58*2 + 0.06*2*width_cm),
         exp_shell3 = exp(0.27+0.04*width_cm - 1.58*3 + 0.06*3*width_cm))

p3 <- data %>% ggplot(aes(x = width_cm)) +
  geom_point(aes(y = satellites_num)) +
  geom_line(aes(y = exp_shell1), color = "black") +
  geom_line(aes(y = exp_shell2), color = "black") +
  geom_line(aes(y = exp_shell3), color = "black") +
  geom_text(aes(x = 35, y = 7.7, label = "Spine 1 (best)"), color="black", show.legend = FALSE) +
  geom_text(aes(x = 34.5, y = 11.9, label = "Spine 2"), color="black", show.legend = FALSE) +
  geom_text(aes(x = 35.1, y = 18.2, label = "Spine 3 (worst)"), color="black", show.legend = FALSE) +
  ylab("Number of Satellites") +
  xlab("Shell Width (cm)") +
  xlim(20, 37) +
  theme_bw()
```

Visualization of Lines

Example:



Visualizations of Lines

Example:

```
quantile(data$width_cm, c(.25, .5, .75))
```

```
##  25%  50%  75%
## 24.9 26.1 27.7
```

```
data <- data %>%
  mutate(exp_width25 = exp(0.27+0.04*25 - 1.58*spine_cond + 0.06*spine_cond*25),
         exp_width26 = exp(0.27+0.04*26 - 1.58*spine_cond + 0.06*spine_cond*26),
         exp_width28 = exp(0.27+0.04*28 - 1.58*spine_cond + 0.06*spine_cond*28))
```

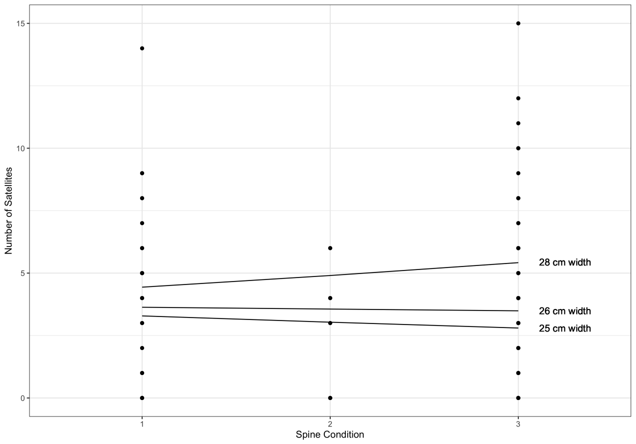
Visualizations of Lines

Example:

```
p4 <- data %>% ggplot(aes(x = spine_cond)) +
  geom_point(aes(y = satellites_num)) +
  geom_line(aes(y = exp_width25), color = "black") +
  geom_line(aes(y = exp_width26), color = "black") +
  geom_line(aes(y = exp_width28), color = "black") +
  geom_text(aes(x = 3.25, y = 5.45, label = "28 cm width"), color="black", show.legend = FALSE) +
  geom_text(aes(x = 3.25, y = 3.5, label = "26 cm width"), color="black", show.legend = FALSE) +
  geom_text(aes(x = 3.25, y = 2.8, label = "25 cm width"), color="black", show.legend = FALSE) +
  ylab("Number of Satellites") +
  scale_x_discrete(name ="Spine Condition", limits=c("1","2","3")) +
  theme_bw()
```

Visualization of Lines

Example:



Inference

Statistical Test for β_i

Hypotheses

$$H_0 : \; \beta_i = \beta_i^{(0)} \mid \beta_i \geq \beta_i^{(0)} \mid \beta_i \leq \beta_i^{(0)}$$

$$H_1 : \; \beta_i \neq \beta_i^{(0)} \mid \beta_i < \beta_i^{(0)} \mid \beta_i > \beta_i^{(0)}$$

Test Statistic

$$z_0 = \frac{\hat{\beta}_i - \beta_i^{(0)}}{\text{SE}_{\hat{\beta}_i}}$$

Rejection Region

Reject H_0 if $p < \alpha$.

Inference

Example:

```
summary(m1)[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
```

The interaction between spine condition and shell width (cm) is significant. This means that the relationship between shell width and number of satellites depends on the spine condition. While number of satellites increases as shell width increases, spine condition matters with smaller shell widths – the best spine condition has the most satellites while the worst spine condition has the fewest satellites.

Inference

Confidence Interval for β_i

$$\hat{\beta}_i \pm z_{1-\alpha/2}\mathrm{SE}_{\hat{\beta}_i}$$

Example:

```
m2 <- glm(satellites_num ~ width_cm + spine_cond,
          family="poisson", data=data)
confint(m2)
```

```
##                  2.5 %      97.5 %
## (Intercept) -4.2711245 -1.95362435
## width_cm     0.1206430  0.20034187
## spine_cond  -0.1430857  0.06075119
```

Inference

We can also find the CI for IRR_i by exponentiating the lower and upper bounds.

Example:

```
round(exp(confint(m2)),3)
```

```
##                2.5 % 97.5 %
## (Intercept) 0.014  0.142
## width_cm    1.128  1.222
## spine_cond  0.867  1.063
```

Predictions

Given a set of values for the predictors in the model, we can return an estimated count.

In linear regression, we returned an expected value.

In logistic regression, we returned a probability.

$$\hat{Y} = \exp\left\{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k\right\}$$

UNIVERSITY of WEST FLORIDA

Predictions

Example:

```
data$p_hat <- predict(m1, type="response")
head(data)
```

```
## # A tibble: 6 x 7
##   color spine_cond width_cm satellites_num weight_g sattelites_yn p_hat
##   <dbl>      <dbl>    <dbl>          <dbl>    <dbl>         <dbl> <dbl>
## 1     3          3     28.3              8     3050             1  3.93
## 2     4          3     22.5              0     1550             0  1.19
## 3     2          1     26                9     2300             1  3.13
## 4     4          3     24.8              0     2100             0  1.91
## 5     4          3     26                4     2600             1  2.45
## 6     3          3     23.8              0     2100             0  1.55
```

Diagnostics

We generally are not worried about residuals in Poisson regression.

We can still look at Cook's distance.

Recall that we look for "spikes" on the graph.

This allows us to determine any leverage/influence points.

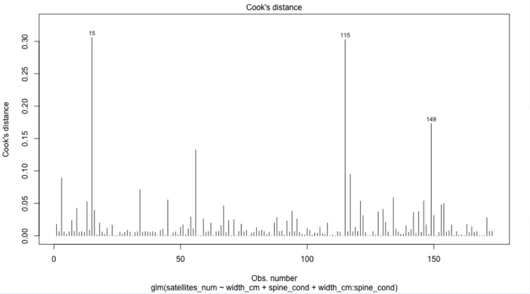
Leverage/influence points are ones that have an effect on the regression model.

If we detect leverage/influence points, we can perform sensitivity analysis to determine how “different” the model is.

Diagnostics

Example:

```
plot(m1, which = 4)
```



Diagnostics

We can also check for multicollinearity using the VIF.

Recall that VIF > 10 indicates multicollinearity.

Example:

```
vif(m1)
```

```
##             width_cm          spine_cond width_cm:spine_cond
##             6.774145          145.636902          145.719376
```

A reminder that we should not include an interaction when checking VIF

Diagnostics

We can also check for multicollinearity using the VIF.

Recall that VIF > 10 indicates multicollinearity.

Example:

```
vif(m2)
```

```
##   width_cm spine_cond
##   1.046582   1.046582
```