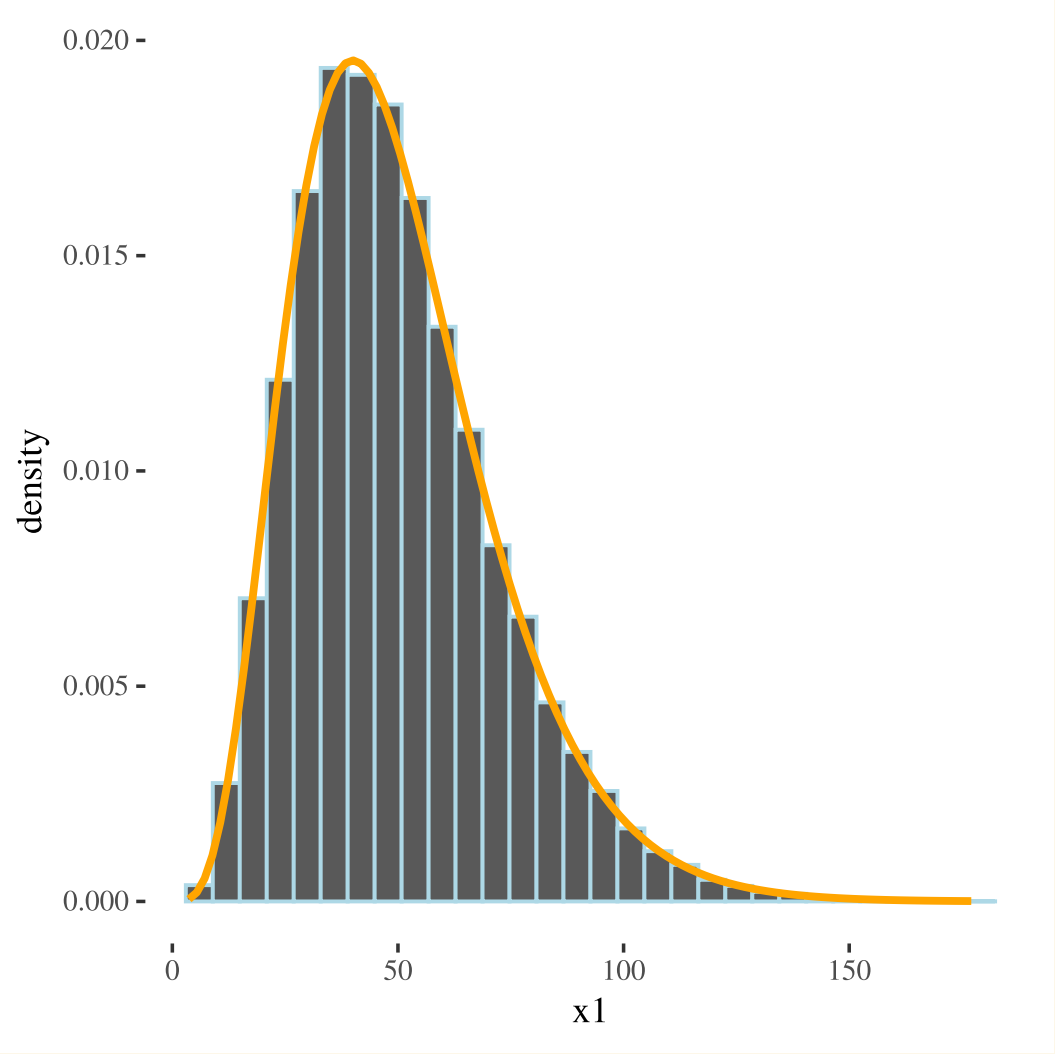
Poisson regression model

Stat 230

May 23 2022

Overview



Today:

Poisson responses

GLM: Poisson regression

EDA

Estimation and inference

Poisson responses

Poisson counts are defined as

- $Y_i =$ number of successes/events that occur in a fixed period of time or region of space

Examples

- Number of possum species per plot
- Number of COVID cases in Rice County in a week
- Number of earthquakes, per year, in the US
- Number of photons emitted per second from an extra-galactic source
- Number of arrests resulting from 911 calls

Poisson response distribution

- If Y has a Poisson distribution, the probability that we observed y successes is

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!} \text{ for any } y = 0, 1, 2, \ldots \text{ and } \mu > 0$$

- Expected number (or mean number) of successes is

$$E(Y) = \mu$$

- SD in the number of successes is

$$SD(Y) = \sqrt{\mu}$$

- If Y has a Poisson distribution, we write

$$Y \sim \text{Poisson}(\mu)$$

Poisson regression assumptions

$$Y \sim \text{Poisson}(\mu)$$

Goal: Model μ as a function of predictors x_1, x_2, $\cdots$ x_p!

Assumption of Poisson regression:

1. **Poisson Response** The response variable is a count per unit of time or space, described by a Poisson distribution.
2. **Independence** The observations must be independent of one another.
3. **Mean=Variance** By definition, the mean of a Poisson random variable must be equal to its variance.
4. **Linearity** The log of the mean rate, log(λ), must be a linear function of x.

Poisson model as a GLM

The kernel mean function defines the expected value (mean $\mu_{y|x}$) of Y as a function of $\eta = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$.

- **MLR Kernel Mean:** $-\infty < \mu < \infty$

$$\mu_{y|x} = \eta = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$

- **Binary Logistic Kernel Mean:** $0 < \pi < 1$

$$\mu_{y|x} = \pi(x) = \frac{e^{\eta}}{1 + e^{\eta}} = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p}}$$

- **Poisson Kernel Mean:** $\mu > 0$

$$\mu_{y|x} = e^{\eta} = e^{\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p}$$

Poisson model as a GLM

The link function defines the linear combination $\eta = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$ as a function of $\mu_{y|x}$.

- **MLR logit (identity) function:** $-\infty < \eta < \infty$

$$\eta = \mu_{y|x} = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$

- **Binary Logit Link function:** $-\infty < \eta < \infty$

$$\eta = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$

- **Poisson "log-linear" link function:** $-\infty < \eta < \infty$

$$\eta = \ln \mu_{y|x} = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$

Poisson Regression model

$$Y_i \mid X_i \overset{\text{indep.}}{\sim} \text{Pois}(\mu(Y_i \mid X_i))$$

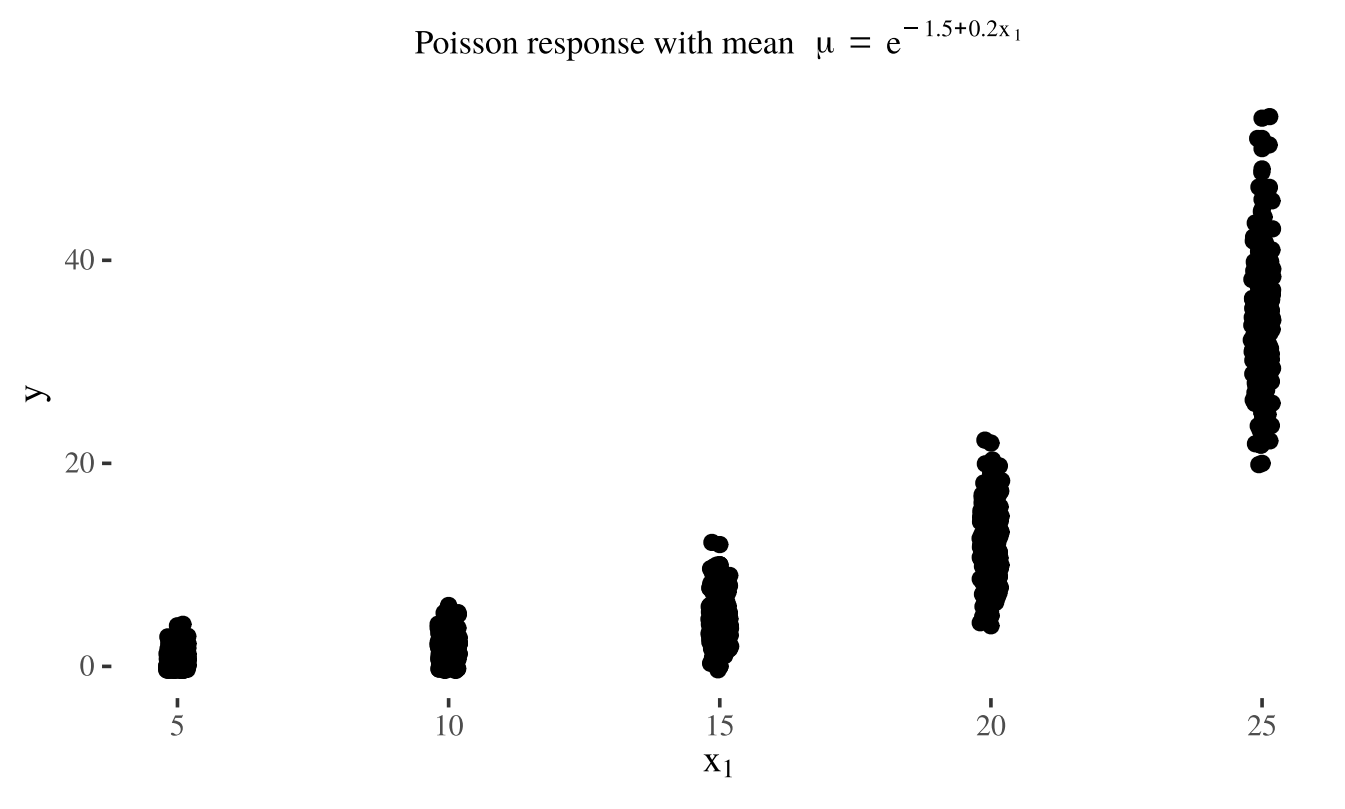
- Changes in x affect the log-mean response ("log-linear" model)

$$\ln(\mu(Y_i \mid X_i)) = \eta_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}$$

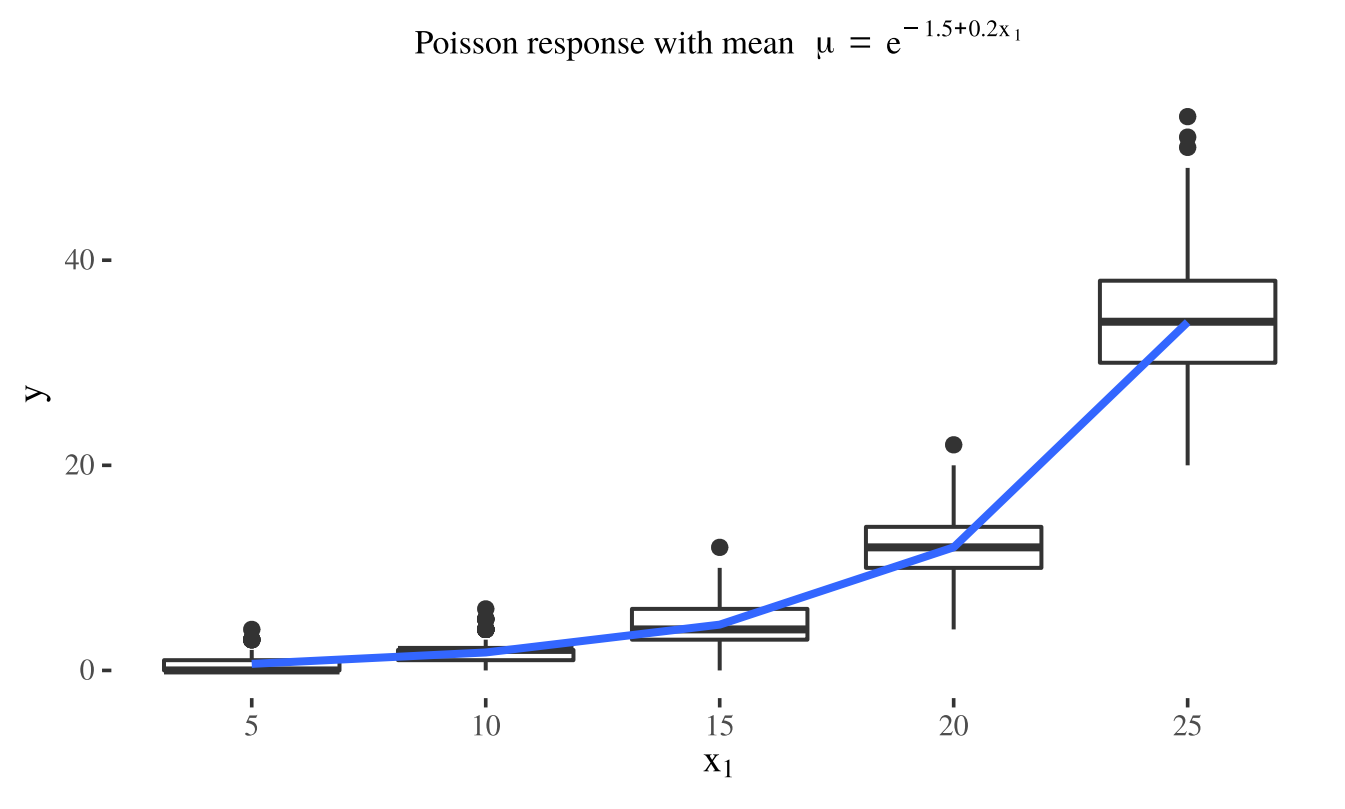
- The mean response is

$$\mu(Y_i \mid X_i) = e^{\eta_i} = e^{\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}}$$

Poisson Regression model example



Poisson Regression model example



Poisson Regression model: estimation

- GLM: estimate β_i using maximum likelihood estimation with likelihood function

$$L(\beta; data) = \prod_{i=1}^{n} \frac{e^{-\mu(X_i)} \mu(X_i)^{y_i}}{y_i!}$$

- MLE theory: $\hat{\beta}_i$ approximately normally distributed when $\mathbf{n}$ is large enough or when $\mu(X_i)$ (estimated means) are large enough.

Poisson Regression model: Inference

Usual Wald (MLE) inference using $N(0,1)$:

- A $C\%$ CI for β : $\hat{\beta} \pm z^* SE(\hat{\beta})$
- Test $H_0 : \beta = 0$ with test stat $z = \frac{\hat{\beta} - 0}{SE(\hat{\beta})}$
- Usual GLM Drop-in-deviance using Poisson (residual) deviance

$$G^2 = 2\left[\ln L\left(\bar{\mu}_i\right) - \ln L\left(\hat{\mu}\left(X_i\right)\right)\right] = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\mu}\left(X_i\right)}\right) - \left(y_i - \hat{\mu}\left(X_i\right)\right)\right]$$

- $L\left(\hat{\mu}\left(X_i\right)\right)$ is the Poisson model likelihood where $\hat{\mu}\left(X_i\right)$ are estimated from the model.
- $L\left(\bar{\mu}_i\right)$ is the saturated model likelihood where $\bar{\mu}_i = y_i$.

Poisson Regression model: Interpretation

$$\mu\left(Y_i \mid X_i\right) = e^{\eta_i} = e^{\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}}$$

Unlogged predictor x_1 :

- A one unit change in x_1 is associated with a e^{β_1} multiplicative change in the mean response.

Logged predictor $\ln(x_1)$:

- A multiplicative change of m in x_1 is associated with a m^{β_1} multiplicative change in the mean response.

Example: Australian Possums

Goal: what factors are associated with good habitat for possums?

- Specifically, how is bark quality associated with possum numbers?
- $n = 1513$-hectare sites in Australia
- $Y =$ number of possum species found on the site
- $X =$ bark quality index, higher values indicate better quality

Why fit a Poisson GLM instead of a SLR? instead of a logistic?

EDA for Poisson counts

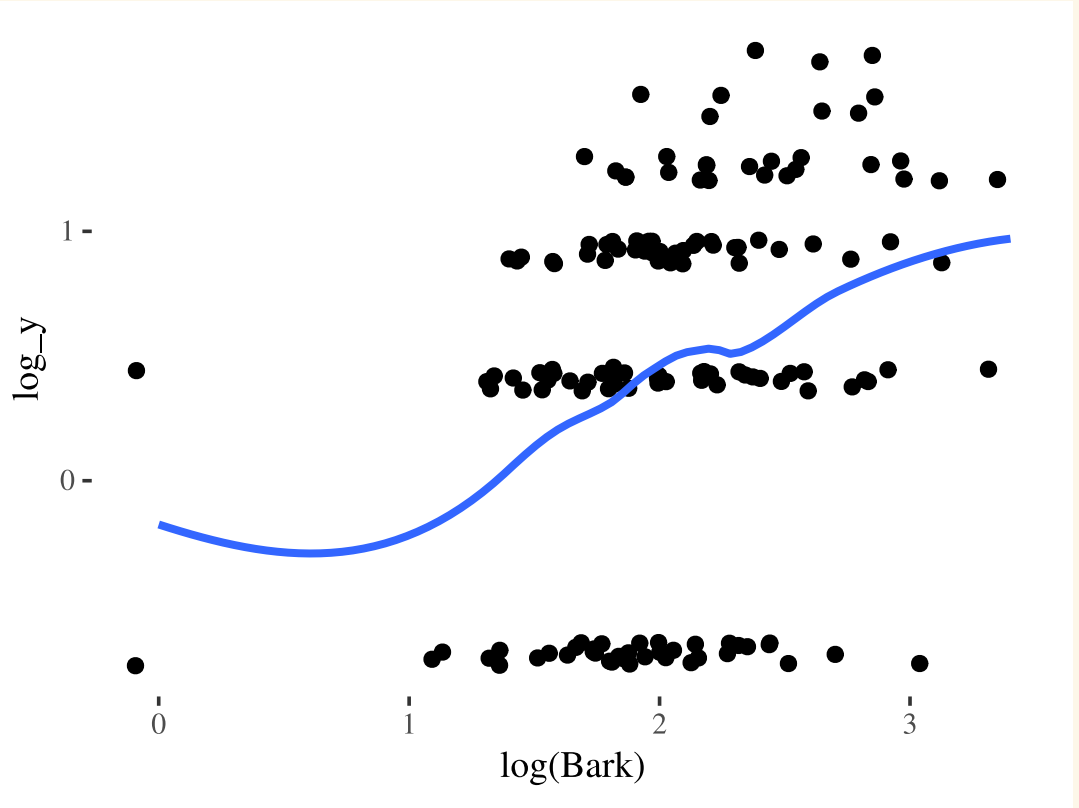
$$\ln(\mu\left(Y_i \mid X_i\right)) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}$$

- EDA: Scatterplot of $\ln(Y)$ or $\ln(Y + 0.5)$ against x should be approximately linear
- Variance does not need to be constant
- If Y contains 0 counts, plot $\ln(Y + 0.5)$ against x

Example: Australian Possums

```
possums <- possums %>%
 mutate(log_y = log(y + 0.5))

ggplot(possums, aes(x = log(Bark), y = log_y)) +
 geom_jitter(height = 0.05, width = 0.1) +
geom_smooth(se = FALSE)
```



Poisson model in R

$\text{glm}(y \sim x1 + x2$, family $=$ poisson, data $=)$

- We don't use log of Y as the response, the Poisson model will convert it to the log mean scale automatically
- Get the fitted values $\hat{\mu} = \hat{y}$ for each case in the data
  - `fitted(my_glm)`
  - `augment(my_glm, type.predict = "response")`
  - `predict(my_glm, type = "response")`

Example: Australian Possums

```
pos_glm <- glm(y ~ log(Bark), # model form
 family = poisson, # poisson model
 data = possums) # data
summary(pos_glm)

Call:
glm(formula = y ~ log(Bark), family = poisson, data = possums)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.18523  -1.26246  -0.07764   0.55078   2.11368

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8801     0.3027  -2.907  0.00365 **
log(Bark)     0.5945     0.1335   4.453 8.45e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 187.49  on 150  degrees of freedom
Residual deviance: 167.51  on 149  degrees of freedom
AIC: 452.31

Number of Fisher Scoring iterations: 5
```

Example: Australian Possums

```
library(broom)
tidy(pos_glm, conf.int = TRUE)
# A tibble: 2 × 7
  term         estimate std.error statistic    p.value conf.low conf.high
  <chr>           <dbl>     <dbl>     <dbl>      <dbl>    <dbl>     <dbl>
1 (Intercept)   -0.880     0.303     -2.91 0.00365       -1.48     -0.295
2 log(Bark)      0.594     0.133      4.45 0.00000845     0.333     0.856
```

- What is the estimated mean number of species as a function of bark index?
- What is the effect of doubling the bark index on the mean number of possum species?
  - assess the stat significant of this effect
  - get a CI for this effect.

Example: Australian Possums

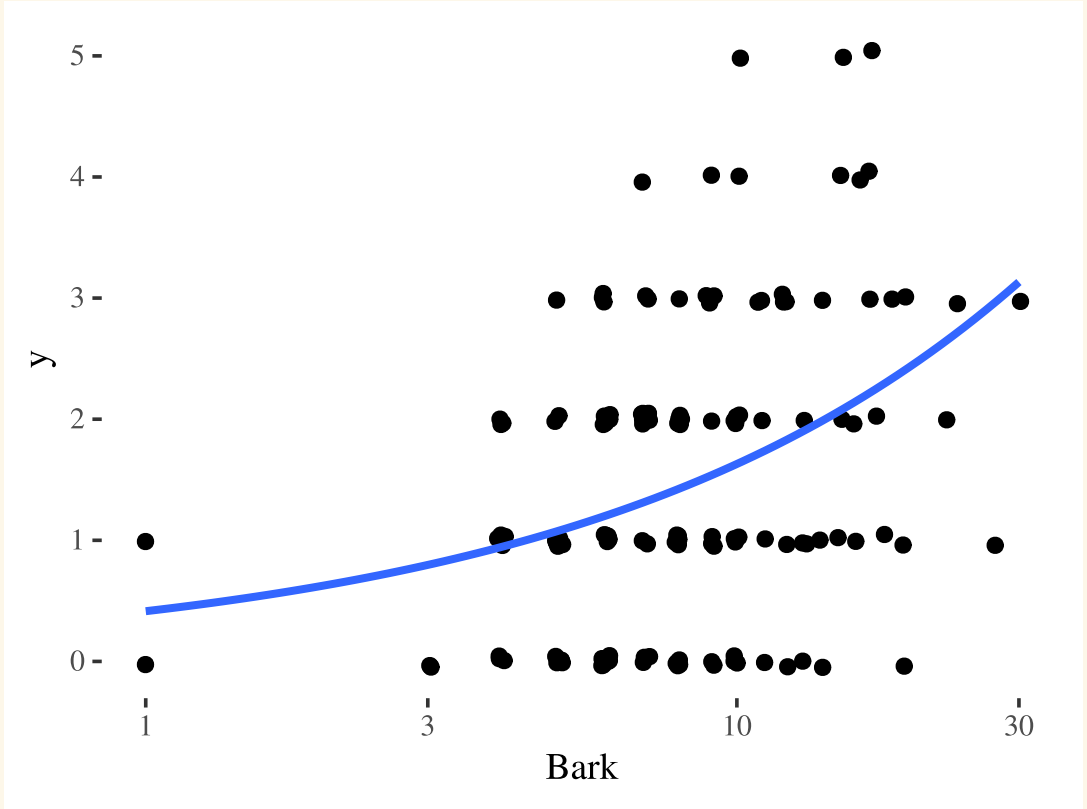
```
tidy(pos_glm, conf.int = TRUE)
# A tibble: 2 × 7
  term         estimate std.error statistic    p.value conf.low conf.high
  <chr>           <dbl>     <dbl>     <dbl>      <dbl>    <dbl>     <dbl>
1 (Intercept)   -0.880      0.303     -2.91 0.00365       -1.48     -0.295
2 log(Bark)      0.594      0.133      4.45 0.00000845     0.333     0.856
```

- What is the estimated mean number of species as a function of bark index?

$$\hat{\mu}(Y \mid x) = e^{-0.880 + 0.594 \ln(\text{ Bark })} = e^{-0.880}\ \text{Bark}^{\ 0.594}$$

Example: Australian Possums

```
ggplot(possums, aes(x = Bark, y = y)) +
 geom_jitter(height = .05) +
  scale_x_log10() +
geom_smooth(method="glm",
            method.args = list(family=poisson),
            se=FALSE)
```

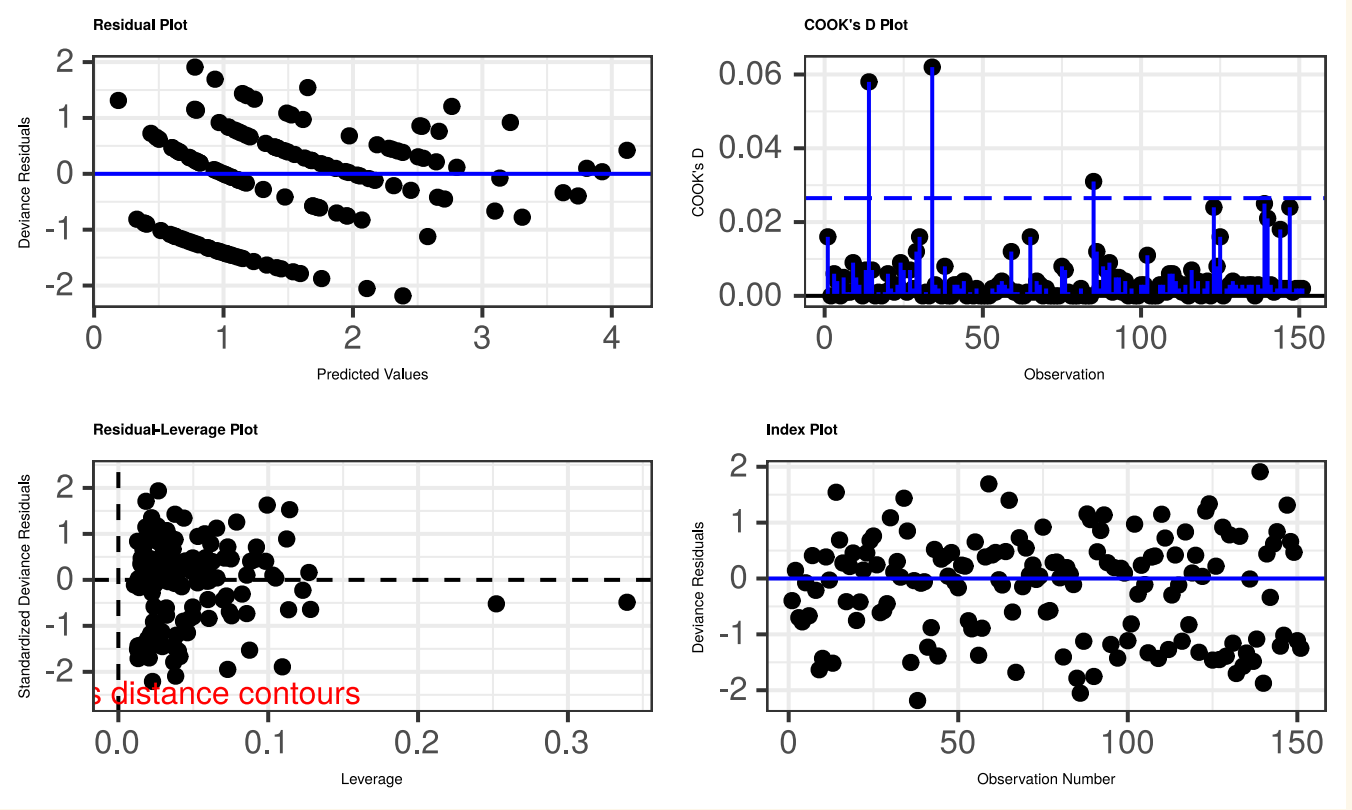


Example: Australian Possums with more predictors

```
possums$stumps <- dplyr::recode(possums$Stumps, `0`="none",`1`="present")
pos_glm_bigger <- glm(y ~ sqrt(Acacia) + log(Bark) + Habitat +
                        log(Shrubs) + log(Stags) + stumps,
                      family=poisson, data=possums)
tidy(pos_glm_bigger)
```

```
# A tibble: 7 × 5
  term           estimate std.error statistic     p.value
  <chr>             <dbl>     <dbl>     <dbl>       <dbl>
1 (Intercept)    -1.98        0.400    -4.95    0.000000730
2 sqrt(Acacia)    0.0893      0.0704    1.27    0.205
3 log(Bark)       0.415       0.161     2.57    0.0101
4 Habitat         0.0588      0.0395    1.49    0.137
5 log(Shrubs)     0.00583     0.118     0.0496  0.960
6 log(Stags)      0.404       0.118     3.41    0.000651
7 stumpspresent  -0.251       0.276    -0.909   0.363
```

Residuals and influence analysis



Residuals and influential analysis is mostly good!

Example: Australian Possums

Can we remove all insignificant terms from the full model?

```
Analysis of Deviance Table

Model 1: y ~ log(Bark) + log(Stags)
Model 2: y ~ sqrt(Acacia) + log(Bark) + Habitat + log(Shrubs) + log(Stags) +
    stumps
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       148     135.61
2       144     124.65  4   10.959  0.02703 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No! At least one removed term is statistically significant (drop-in-deviance = 10.96, df = 4, p-value = 0.027)

Example: Australian Possums

```
car::vif(pos_glm_bigger)
sqrt(Acacia)    log(Bark)      Habitat  log(Shrubs)   log(Stags)       stumps
    1.770968     1.357957     2.262461     1.356994     1.655739     1.064056
```

Add `habitat` back in (some collinearity with other terms)

```
pos_glm_red <- glm(y ~ log(Bark) + Habitat + log(Stags), family="poisson", data = possums)
anova(pos_glm_red, pos_glm_bigger, test="Chisq")
Analysis of Deviance Table

Model 1: y ~ log(Bark) + Habitat + log(Stags)
Model 2: y ~ sqrt(Acacia) + log(Bark) + Habitat + log(Shrubs) + log(Stags) +
    stumps
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       147     127.09
2       144     124.65  3   2.4463   0.4851
```

We can remove acacia (area of acacia at the site), shrubs (number of shrubs), and stumps (presence of stumps from logging)

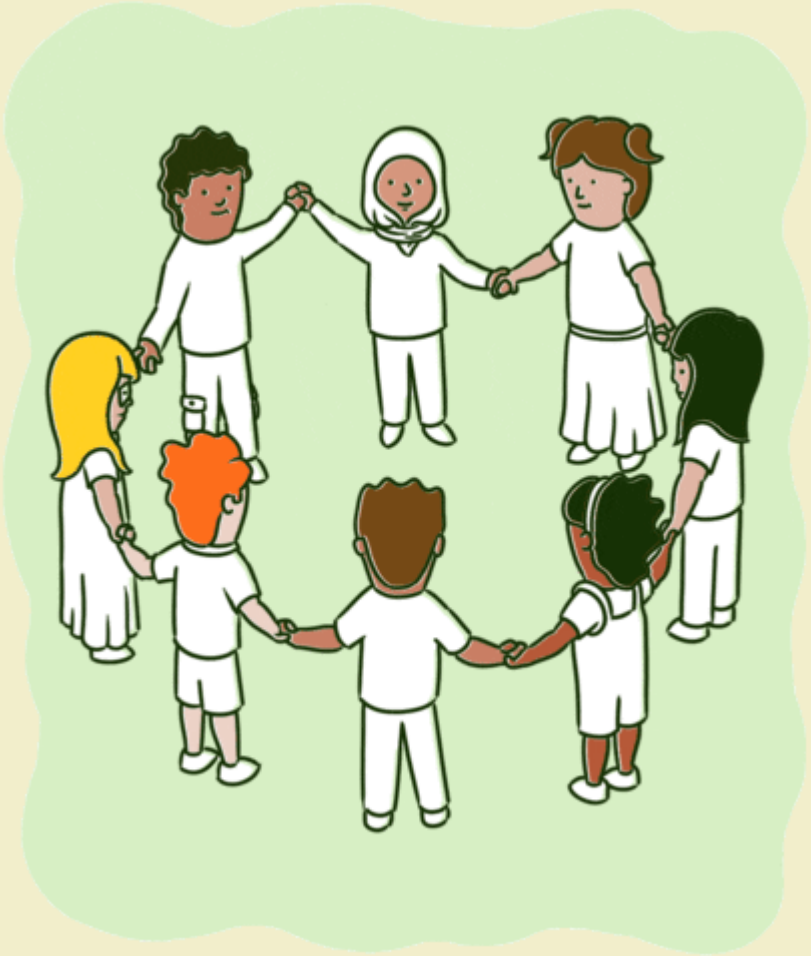
Example: Australian Possums

The statistically significant terms are bark quality, habitat score (higher is better), and stags (number of hollow trees)

```
tidy(pos_glm_red)
# A tibble: 4 × 5
  term         estimate std.error statistic     p.value
  <chr>           <dbl>     <dbl>     <dbl>       <dbl>
1 (Intercept)  -1.75       0.344      -5.09 0.000000358
2 log(Bark)     0.394      0.139       2.83 0.00460
3 Habitat       0.0877     0.0307      2.86 0.00425
4 log(Stags)    0.370      0.109       3.40 0.000675
```

Your Turn 1

05:00



- Go over to the in class activity file
- Go over the class activity in your group