

GLM for count data: Poisson regression

Types of response variable

- ▶ Gaussian: `lm`

Types of response variable

- ▶ Gaussian: `lm`
- ▶ Bernoulli / Binomial: `glm (family binomial / quasibinomial)`

Types of response variable

- ▶ Gaussian: `lm`
- ▶ Bernoulli / Binomial: `glm (family binomial / quasibinomial)`
- ▶ Counts: `glm (family poisson / quasipoisson)`

Poisson regression

- ▶ Response variable: Counts (0, 1, 2, 3...) - discrete

Then

$$\log(N) = a + bx$$

$$N = e^{a+bx}$$

Poisson regression

- ▶ Response variable: Counts (0, 1, 2, 3...) - discrete
- ▶ Link function: \log

Then

$$\log(N) = a + bx$$

$$N = e^{a+bx}$$

Example dataset: Seedling counts in quadrats

```
seedl <- read.csv("data-raw/seedlings.csv")
```

X	count	row	col
Min. : 1.00	Min. :0.00	Min. :1	Min. : 1.0
1st Qu.:13.25	1st Qu.:1.00	1st Qu.:2	1st Qu.: 3.0
Median :25.50	Median :2.00	Median :3	Median : 5.5
Mean :25.50	Mean :2.14	Mean :3	Mean : 5.5
3rd Qu.:37.75	3rd Qu.:3.00	3rd Qu.:4	3rd Qu.: 8.0
Max. :50.00	Max. :7.00	Max. :5	Max. :10.0

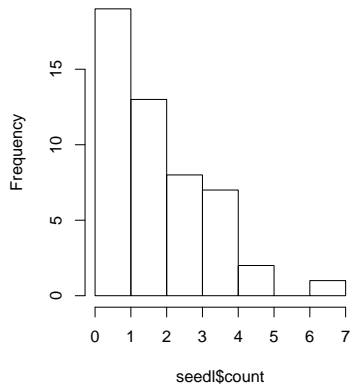
light	area
Min. : 2.571	Min. :0.25
1st Qu.:26.879	1st Qu.:0.25
Median :47.493	Median :0.50
Mean :47.959	Mean :0.62
3rd Qu.:67.522	3rd Qu.:1.00
Max. :99.135	Max. :1.00

EDA

```
table(seed1$count)
```

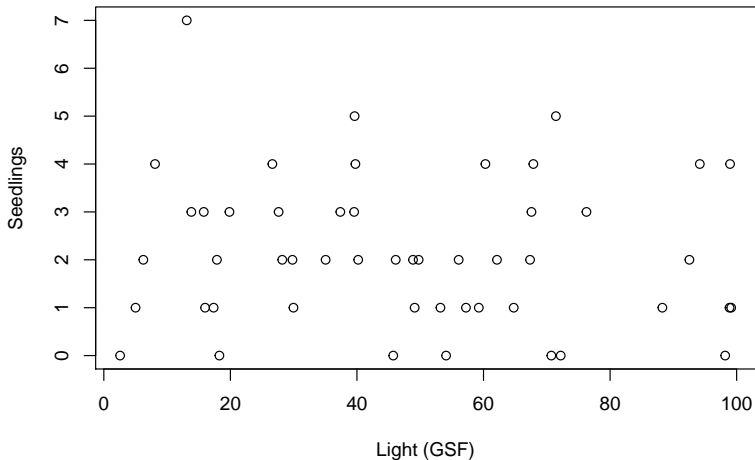
0	1	2	3	4	5	7
7	12	13	8	7	2	1

Histogram of seed1\$count



Q: Relationship between Nseedlings and light?

```
plot(seedl$light, seedl$count, xlab = "Light (GSF)", ylab = "Seedlings")
```



Let's fit model (Poisson regression)

```
seed1.glm <- glm(count ~ light, data = seed1, family = poisson)
summary(seed1.glm)
```

Call:

```
glm(formula = count ~ light, family = poisson, data = seed1)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.1906	-0.8466	-0.1110	0.5220	2.4577

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.881805	0.188892	4.668	3.04e-06 ***
light	-0.002576	0.003528	-0.730	0.465

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 63.029 on 49 degrees of freedom
Residual deviance: 62.492 on 48 degrees of freedom
AIC: 182.03

Number of Fisher Scoring iterations: 5

Interpreting Poisson regression output

Parameter estimates (log scale):

```
coef(seed1.glm)
```

(Intercept)	light
0.881805022	-0.002575656

We need to back-transform: apply the inverse of the logarithm

```
exp(coef(seed1.glm))
```

(Intercept)	light
2.4152554	0.9974277

Using effects package

```
summary(allEffects(seed1.glm))
```

```
model: count ~ light
```

```
light effect
```

```
light
```

	3	30	50	70	100
	2.396665	2.235657	2.123408	2.016794	1.866826

```
Lower 95 Percent Confidence Limits
```

```
light
```

	3	30	50	70	100
	1.684579	1.795202	1.753373	1.567785	1.228247

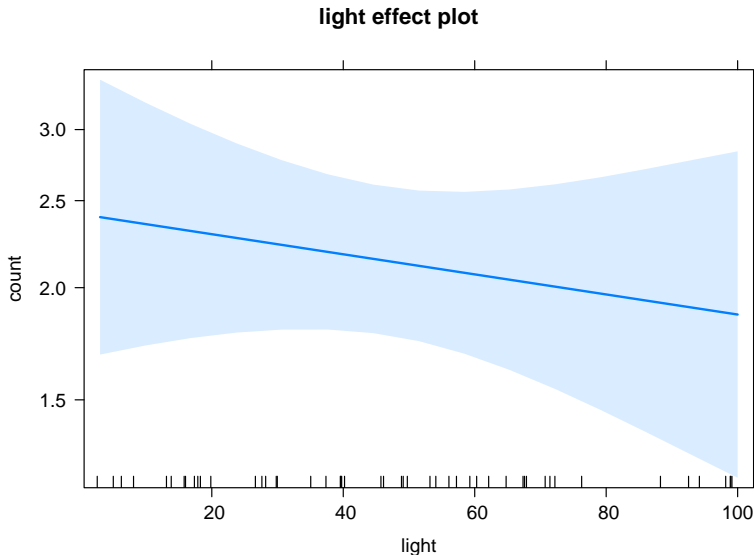
```
Upper 95 Percent Confidence Limits
```

```
light
```

	3	30	50	70	100
	3.409754	2.784179	2.571535	2.594398	2.837408

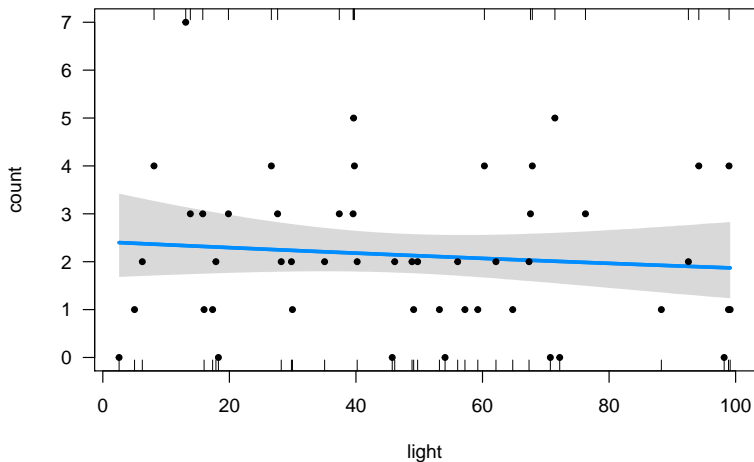
So what's the relationship between Nseedlings and light?

```
plot(allEffects(seed1.glm))
```



Using visreg

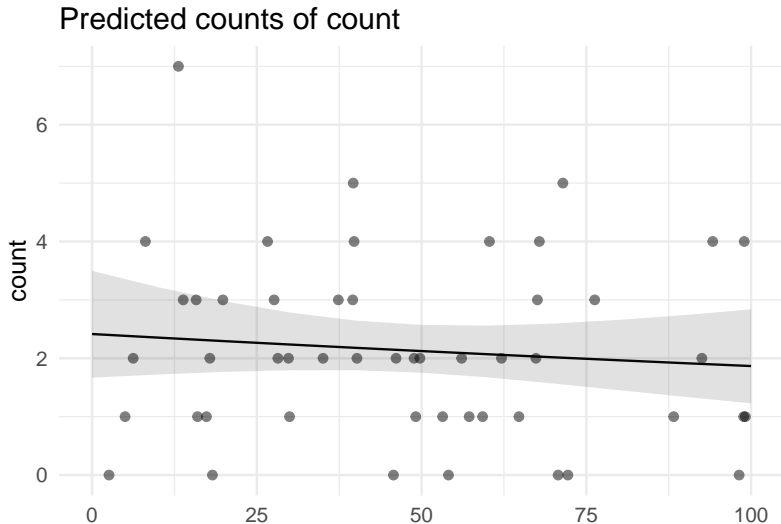
```
visreg(seedl.glm, scale = "response", ylim = c(0, 7))  
points(count ~ light, data = seedl, pch = 20)
```



Using sjPlot

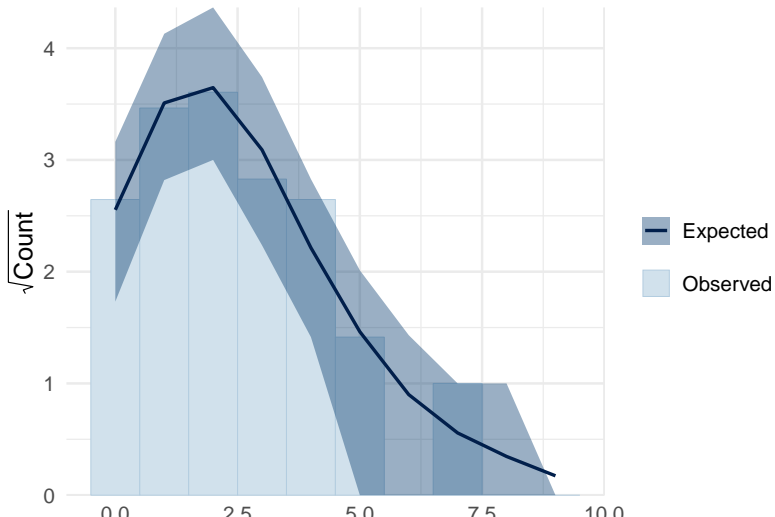
```
sjPlot::plot_model(seed1.glm, type = "eff", show.data = TRUE)
```

\$light

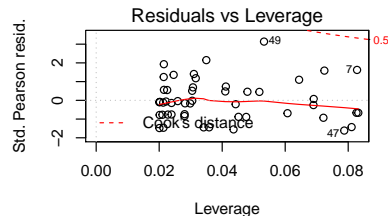
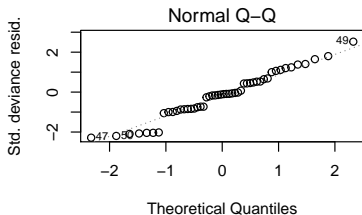
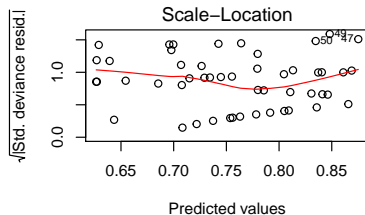
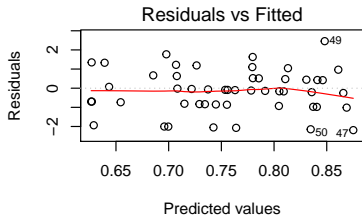


Calibration plot with count data: rootograms

```
sims <- simulate(seedl.glm, nsim = 100)
yrep <- t(as.matrix(sims))
bayesplot::ppc_rootogram(seedl$count, yrep)
```



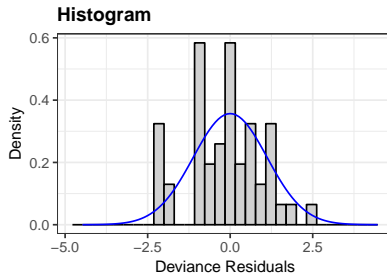
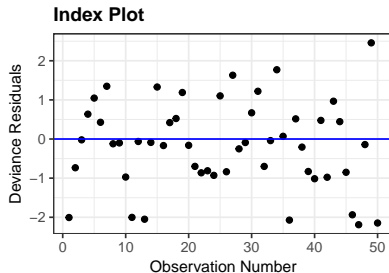
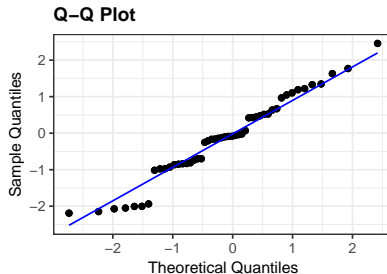
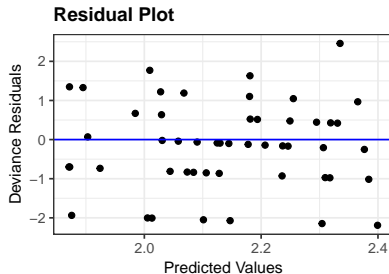
Poisson regression: model checking



null device

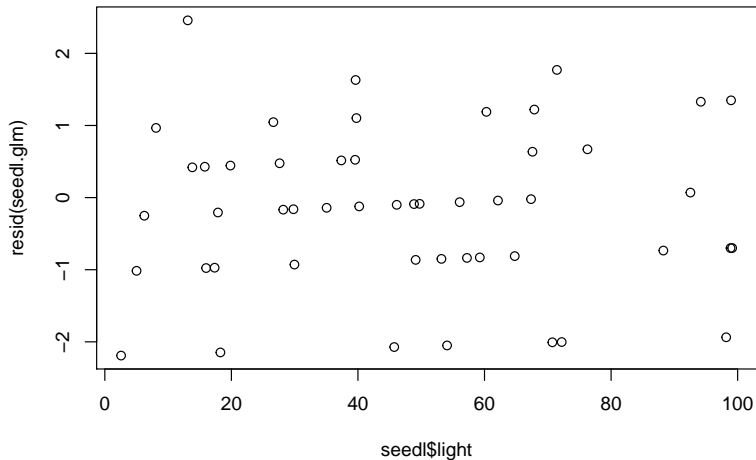
Poisson regression: model checking

```
ggResidpanel::resid_panel(seed1.glm)
```



Is there pattern of residuals along predictor?

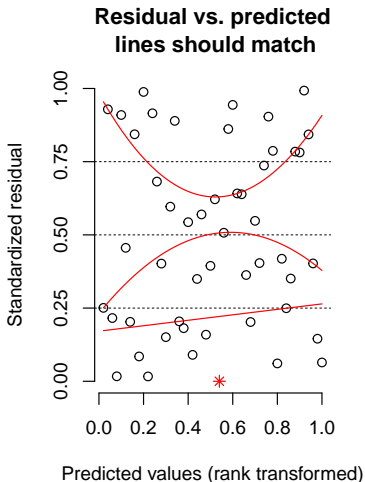
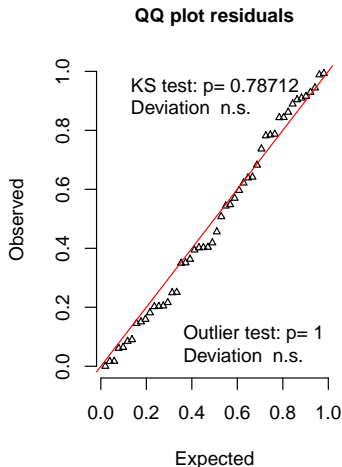
```
plot(seed1$light, resid(seed1.glm))
```



Residuals diagnostics with DHARMA

```
DHARMA::simulateResiduals(seed1.glm, plot = TRUE)
```

DHARMA scaled residual plots



Poisson regression: Overdispersion

Always check overdispersion with count data

```
simres <- simulateResiduals(seed1.glm, refit = TRUE)  
testDispersion(simres, plot = FALSE)
```

DHARMA nonparametric dispersion test via mean deviance residuals
fitted vs. simulated-refitted

```
data:  simres  
dispersion = 1.1655, p-value = 0.432  
alternative hypothesis: two.sided
```

Accounting for overdispersion in count data

Use family quasipoisson

Call:

```
glm(formula = count ~ light, family = quasipoisson, data = seedl
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.1906	-0.8466	-0.1110	0.5220	2.4577

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.881805	0.201230	4.382	6.37e-05 ***
light	-0.002576	0.003758	-0.685	0.496

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.1349

Null deviance:	63.029	on 49	degrees of freedom
Residual deviance:	62.492	on 48	degrees of freedom

Mean estimates do not change after accounting for overdispersion

```
model: count ~ light
```

```
light effect
```

```
light
```

	3	30	50	70	100
	2.396665	2.235657	2.123408	2.016794	1.866826

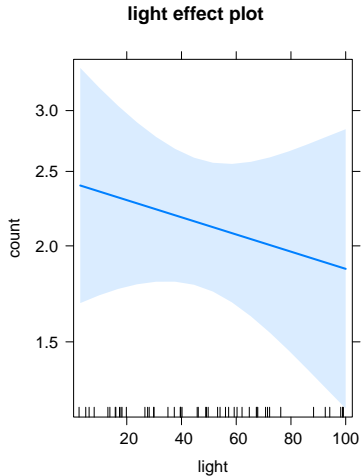
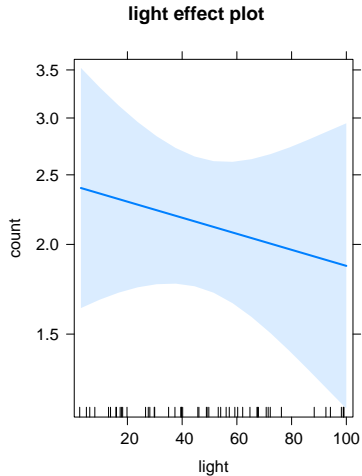
```
model: count ~ light
```

```
light effect
```

```
light
```

	3	30	50	70	100
	2.396665	2.235657	2.123408	2.016794	1.866826

But standard errors may change



What if survey plots have different area?

Avoid regression of ratios

seedlings/area \sim light

J. R. Statist. Soc. A (1993)
156, Part 3, pp. 379–392

Spurious Correlation and the Fallacy of the Ratio Standard Revisited

By RICHARD A. KRONMAL†

Figure 1:

Use offset to standardise response variables in GLMs

```
seedl.offset <- glm(count ~ light, offset = seedl$area, data = seedl, family = poisson)
summary(seedl.offset)
```

Call:

```
glm(formula = count ~ light, family = poisson, data = seedl,
     offset = seedl$area)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6926	-0.8532	0.1491	0.5211	3.1051

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.299469	0.185468	1.615	0.106
light	-0.004498	0.003441	-1.307	0.191

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 70.263 on 49 degrees of freedom

Note estimates now referred to area units

```
exp(coef(seedl.offset))
```

(Intercept)	light
1.3491422	0.9955123

Other examples

- ▶ Infant mortality \sim GDP

Other examples

- ▶ Infant mortality \sim GDP
- ▶ Number of cones consumed by squirrels (data)