# GLMs for count data: Poisson errors

**Practical 3**

David Orme

The lecture introduced the basic concept of generalised linear models (GLM): we identify a link function that gives a good scale for fitting linear models (the *linear predictor*) and evaluate the fit of that model on the original data using an appropriate statistical distribution (the *error structure*).

In practice, this is very similar to using a linear model – the key thing to be aware of is when we're using the scale of the linear predictor and when we're using the scale of the original data. The practical will give a walk through with one dataset and then provide examples to try.

## Species richness on the Galapagos Islands

This dataset is on the number of plant species found on the Galapagos Islands[1]. It records the total number and number of endemic species along with information on the size and maximum elevation of the island and position in the archipelago. For more information, see the R package faraway.

We're going to use it for a very simple GLM analysis – is there a relationship between the area of the island and the number of plant species found there?

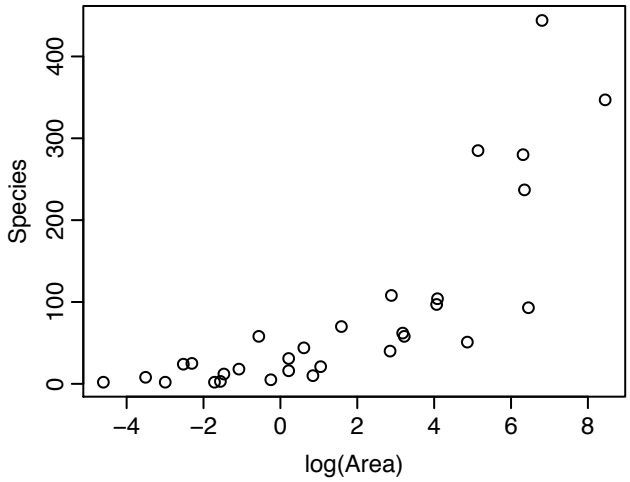
```
odonata <- read.delim('gala.txt')
```

```
str(gala)
```

```
## 'data.frame': 30 obs. of  7 variables:
##  $ Species  : int  58 31 3 25 2 18 24 10 8 2 ...
##  $ Endemics : int  23 21 3 9 1 11 0 7 4 2 ...
##  $ Area     : num  25.09 1.24 0.21 0.1 0.05 ...
##  $ Elevation: int  346 109 114 46 77 119 93 168 71 112 ...
##  $ Nearest  : num  0.6 0.6 2.8 1.9 1.9 8 6 34.1 0.4 2.6 ...
##  $ Scruz    : num  0.6 26.3 58.7 47.4 1.9 ...
##  $ Adjacent : num  1.84 572.33 0.78 0.18 903.82 ...
```

```
plot(Species ~ log(Area) , data=gala)
```

[1] M. P. Johnson and P. H. Raven (1973) 'Species number and endemism: The Galapagos Archipelago revisited' Science, 179, 893-895




We can see from the plot that there appears to be a very strong relationship. The problem is that the data is count data: there is increasing variance and the data is bounded below at zero. In October, we analysed similar data with a log transformation – now we're going to do it properly!

In order to fit a GLM, the only thing we need to change is to use the **glm**() function and specify the error distribution using the family option. This option also sets the *link function* to be used. For Poission data, the log link is the default, so we can just say family=poisson but to be clear we can say family=**poisson**(link=log).

```
gala$lgArea    <- log(gala$Area)
galaMod <- glm(Species ~ lgArea, data=gala, family=poisson(link=log))
```

That's it! Just as with a linear model we can look at the significance of the model terms and coefficients. The only difference is that the model terms are tested using changes in model deviance not variance, so although we use the same anova function, we're doing *analysis of deviance*. For a Poisson GLM, we use a $\chi^2$ test rather than an $F$ test but otherwise we're doing the same assessment of how much deviance a term explains.

```
anova(galaMod, test='Chisq')

## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Species
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                      29       3511
## lgArea  1     2859        28        652   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It doesn't label it as such, but the change in deviance is used as the $\chi^2$ value, with the degrees of freedom equal to the number of degrees of freedom used to fit the term. We can calculate it below, but R replaces the zero with <2e-16, which essentially means smaller than can be accurately calculated.

```
1 - pchisq(2859, df=1)
```

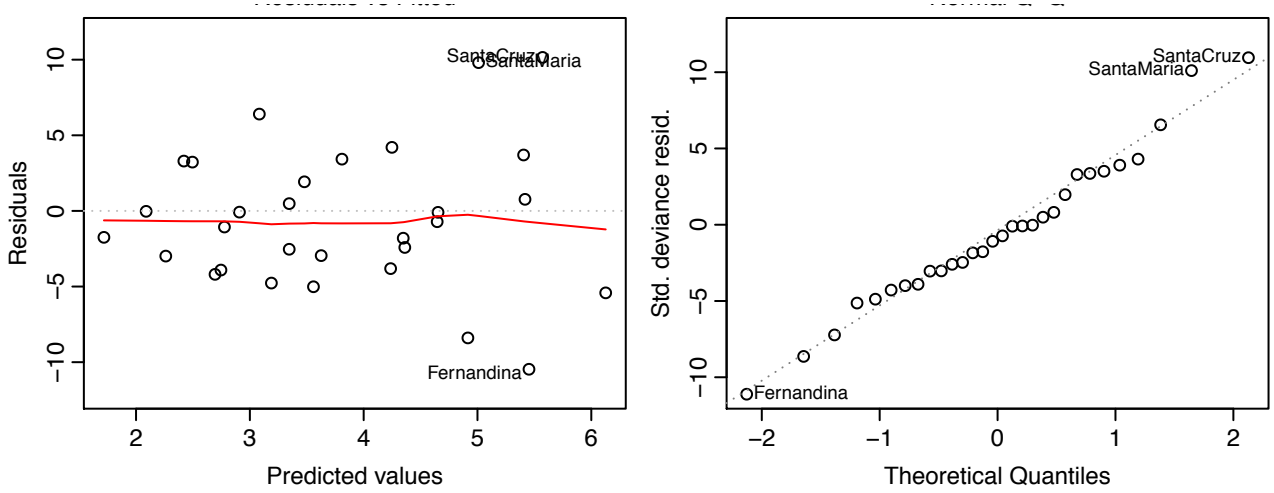
```
## [1] 0
```

The summary of the coefficients from the model is very similar to the linear model output but doesn't include $r^2$. This can't be defined for a GLM, since the residual sums of squares don't make sense as a measure of model fit, but we can calculate the proportion of the null deviance explained, which does a similar job.

```
summary(galaMod)

##
## Call:
## glm(formula = Species ~ lgArea, family = poisson(link = log),
##     data = gala)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -10.469   -3.607   -0.887    2.903   10.152
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.27320    0.04166    78.6   <2e-16 ***
## lgArea       0.33774    0.00715    47.2   <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: 3510.73  on 29  degrees of freedom
## Residual deviance:  651.67  on 28  degrees of freedom
## AIC: 816.5
##
## Number of Fisher Scoring iterations: 5

(galaMod$null.deviance - galaMod$deviance)/galaMod$null.deviance

## [1] 0.814
```

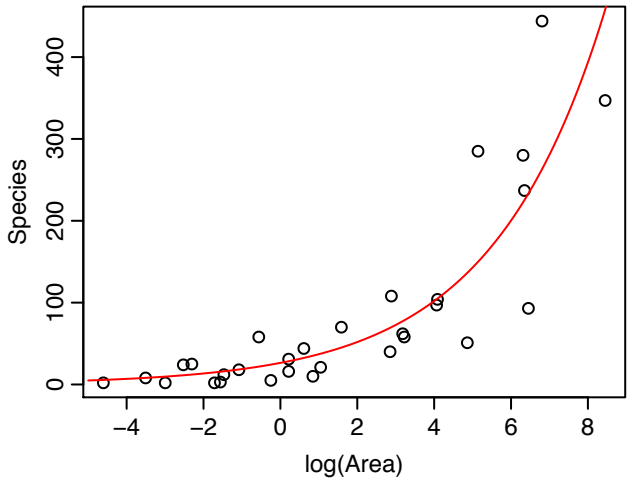
The same model fitting tools such as **drop1**(), **update**() and **step**() also still work - try them out! We still should examine the diagnostic plots – these plots now use the deviance residuals, and should be still be normally distributed with constant variance.

```
par(mfrow=c(1,2))
plot(galaMod, which=c(1,2))
```




The bit that can be confusing is in plotting the model. The coefficients are *on the scale of the linear predictor*, so plotting them over the data, which we never actually transform and which might have zeros, needs some attention. The easy approach for the actual fitted means in the model is to get **predict**() to give us the predictions on the scale of the response.

```
# predict for a neat sequence of log area values
pred <- expand.grid(lgArea = seq(-5, 9, by=0.1))
pred$fit <- predict(galaMod, newdata=pred, type='response')

# plot the logged data and the model lines
plot(Species ~ log(Area), data=gala)
lines(fit ~ lgArea, data=pred, col='red')
```

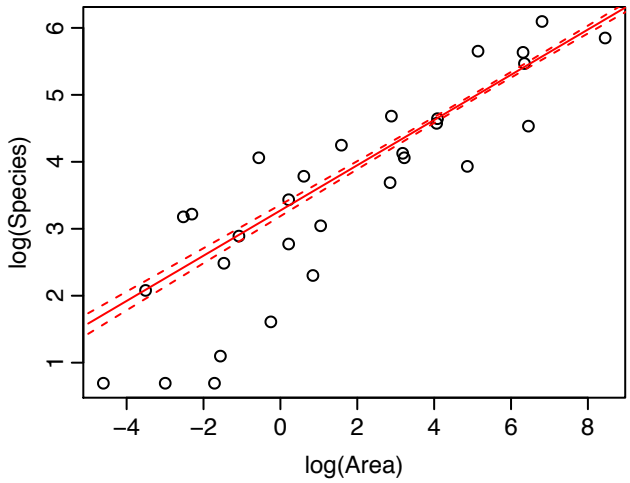



However, if we want to show confidence limits, then this isn't so simple as we can't just multiply the standard error by the critical $t$ value on the count scale. Instead we first approximate confidence limits on the scale of the linear predictor[2]:

```
# predict for a neat sequence of log area values
pred <- expand.grid(lgArea = seq(-5, 9, by=0.1))
predMod <- predict(galaMod, newdata=pred, se.fit=TRUE)

# get the fit and confidence limits
pred$fit <- predMod$fit
```

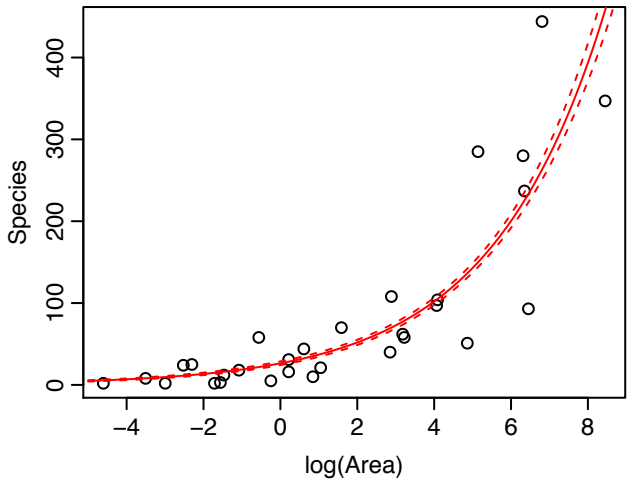
[2]You may be suspicious about how narrow these confidence limits are – you'd be right. We'll come back to this later.

```
pred$se.fit <- predMod$sefit
pred$confint <- predMod$se.fit * qt(0.975, df=28)

# plot the logged data and the model lines
plot(log(Species) ~ log(Area), data=gala)
lines(fit ~ lgArea, data=pred, col='red')
lines(fit + confint  ~ lgArea, data=pred, col='red', lty=2)
lines(fit - confint  ~ lgArea, data=pred, col='red', lty=2)
```




We can now back transform them on to the data. This means we need to know the inverse link function for our model, which for **log**() is **exp**().

```
# plot the data and transformed model lines
plot(Species ~ log(Area), data=gala)
lines(exp(fit) ~ lgArea, data=pred, col='red')
lines(exp(fit + confint)  ~ lgArea, data=pred, col='red', lty=2)
lines(exp(fit - confint)  ~ lgArea, data=pred, col='red', lty=2)
```




## Amphibian roadkills in Portugal

The next dataset fo you to try shows counts of the number of dead amphibians in 500 metre sections of a road in Portugal. There are a huge number of variables in the data frame measuring the local habitat characteristics. We won't use them all in the example.

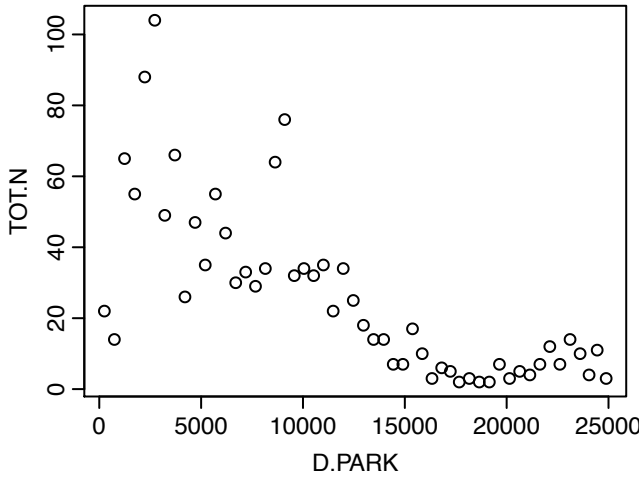
```
roadkill <- read.delim('RoadKills.txt')
```

```
str(roadkill)
```

```
## 'data.frame': 52 obs. of  23 variables:
##  $ Sector      : int  1 2 3 4 5 6 7 8 9 10 ...
## $ X         : int  260181 259914 259672 259454 259307 259189 259092 258993 258880 258767 ...
## $ Y         : int  256546 256124 255688 255238 254763 254277 253786 253296 252809 252322 ...
##  $ BufoCalamita: int  5 1 40 27 67 56 27 37 8 16 ...
##  $ TOT.N       : int  22 14 65 55 88 104 49 66 26 47 ...
##  $ S.RICH      : int  3 4 6 5 4 7 7 7 7 6 ...
##  $ OPEN.L      : num  22.7 24.7 30.1 50.3 43.6 ...
##  $ OLIVE       : num  60.3 40.8 23.7 14.9 35.4 ...
##  $ MONT.S      : num  0 0 0.258 1.783 2.431 ...
##  $ MONT        : num  0.653 0.161 10.918 26.454 11.33 ...
##  $ POLIC       : num  4.811 2.224 1.946 0.625 0.791 ...
##  $ SHRUB       : num  0.406 0.735 0.474 0.607 0.173 ...
##  $ URBAN       : num  7.787 27.15 28.086 0.831 2.452 ...
##  $ WAT.RES     : num  0.043 0.182 0.453 0.026 0 0.039 0.114 0.224 0.177 0 ...
##  $ L.WAT.C     : num  0.583 1.419 2.005 1.924 2.167 ...
##  $ L.D.ROAD    : num  3330 2587 2150 4223 2219 ...
##  $ L.P.ROAD    : num  1.975 1.761 1.25 0.666 0.653 ...
##  $ D.WAT.RES   : num  252.1 139.6 59.2 277.8 967.8 ...
##  $ D.WAT.COUR  : num  735 134.1 269 48.8 126.1 ...
##  $ D.PARK      : num  250 741 1240 1740 2232 ...
##  $ N.PATCH     : num  122 96 67 63 59 49 35 55 52 26 ...
##  $ P.EDGE      : num  554 457 432 421 408 ...
##  $ L.SDI       : num  1.8 1.89 1.93 1.86 1.82 ...
```

Fit a Poisson GLM that predicts the number of road kills (TOT.N) as a function of distance to a nearby natural park (D.PARK).

```
plot(TOT.N ~ D.PARK, data = roadkill)
```



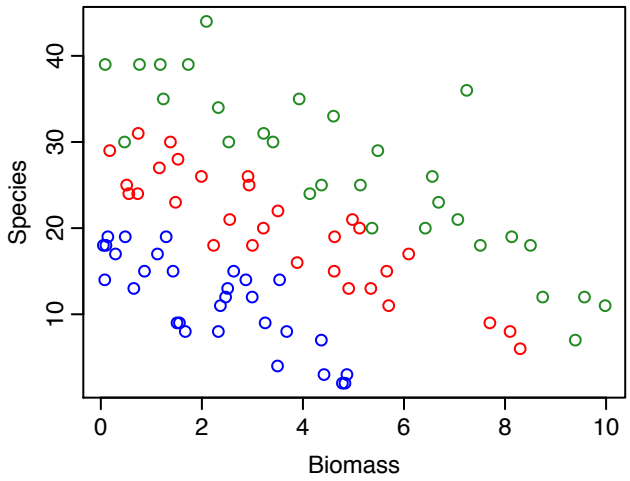

## Species richness in grassland plot

A third dataset includes records of plant species richness from 90 agricultural plots with differing soil pH (a three-level factor) and biomass (a continuous variable). Use this dataset to model whether species richness is

predicted by soil pH and biomass and their interaction.

```
species <- read.delim('species.txt')
```

```
str(species)

## 'data.frame': 90 obs. of  3 variables:
##  $ pH     : Factor w/ 3 levels "high","low","mid": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Biomass: num  0.469 1.731 2.09 3.926 4.367 ...
##  $ Species: int  30 39 44 35 25 29 23 18 19 12 ...

plot(Species ~ Biomass, data=species, col=pH)
```



# Report statistics

## Species richness on the Galapagos

### Methods

I fitted a Poisson GLM with a log link to predict plant species richness on islands of the Galapagos as a function of island area in km$^2$. I tested model significance using $\chi^2$ tests and assessed model suitability using diagnostic plots of deviance residuals.

### Results

Log island area is highly explanatory of species richness ($\chi^2_1 = 2859, p << 0.0001$) and the model explains $81.4\%$ of the deviance in plant species richness. The model is shown in Figure 1 and the model coefficients are shown in Table 1

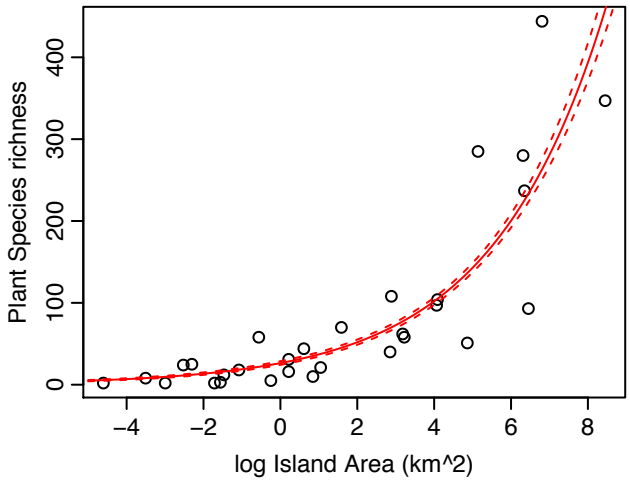



**Figure 1:** *Plant species richness on island of the Galapagos as a function of island area. Predictions of a linear model (solid line) are shown with 95% confidence limits (dashed lines).*

Table 1: Model coefficients, standard errors and signficance tests for a generalised linear model with Poisson errors of plant species richness as a function of island area in the Galapagos.

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 3.2732 | 0.0417 | 78.56 | <0.0001 |
| lgArea | 0.3377 | 0.0072 | 47.21 | <0.0001 |

## Amphibian roadkills in Portugal

### Methods

I fitted a Poisson GLM with a log link to predict the number of amphibian roadkills observed within 500 metre sections of road as a function of distance of the road segment from a National Park in kilometres. I tested model significance using $\chi^2$ tests and assessed model suitability using diagnostic plots of deviance residuals.

### Results

Proximity to a national park is highly explanatory of species richness ($\chi^2_1 = 681, p << 0.0001$) and the model explains $63.5\%$ of the deviance in number of roadkills. The model is shown in Figure 2 and the model coeffi-

cients are shown in Table 2.

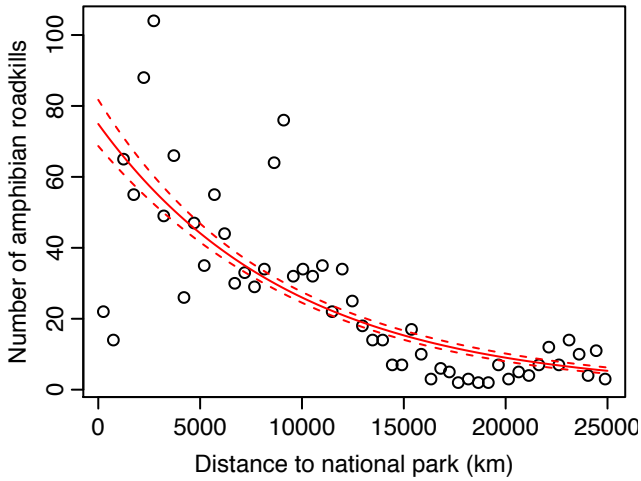



**Figure 2:** *Amphibian roadkills as a function of proximity to a National Park. Predictions of a generalised linear model (solid line) are shown with 95% confidence limits (dashed lines).*

Table 2: Model coefficients, standard errors and signficance tests for a generalised linear model with Poisson errors of amphibian roadkill numbers as a function of proximity to a national park

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 4.3165 | 0.0432 | 99.87 | <0.0001 |
| D.PARK | -0.0001 | 0.0000 | -24.13 | <0.0001 |

## Species richness in grassland plots

### Methods

I fitted a Poisson GLM with a log link to predict the number of species within grassland plots as a function of plant biomass in the plot, soil pH (a three level factor of high, medium and low pH) and their interaction. I tested model significance using $\chi^2$ tests and assessed model suitability using diagnostic plots of deviance residuals.

### Results

All three terms in the model are strongly explanatory of plant species richness: biomass ($\chi^2_1 = 44.7, p << 0.0001$); pH ($\chi^2_2 = 308.4, p << 0.0001$); and the interaction between biomass and pH ($\chi^2_2 = 16.0, p = 0.00033$). The model explains $81.6\%$ of the deviance in the data. The coefficient of the model are shown in Table 3 and the model is shown in Figure 3: plant species richness declines sharply with increasing biomass. Species richness at low biomass is highest and declines faster in higher pH soils.

Table 3: Model coefficients, standard errors and signficance tests for a generalised linear model with Poisson errors of plant species richness in grassland plots as a function of plant biomass and soil pH.

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 3.7681 | 0.0615 | 61.24 | <0.0001 |
| Biomass | -0.1071 | 0.0125 | -8.58 | <0.0001 |
| pHlow | -0.8156 | 0.1028 | -7.93 | <0.0001 |
| pHmid | -0.3315 | 0.0922 | -3.60 | 0.0003 |
| Biomass:pHlow | -0.1550 | 0.0400 | -3.87 | 0.0001 |
| Biomass:pHmid | -0.0319 | 0.0231 | -1.38 | 0.1670 |

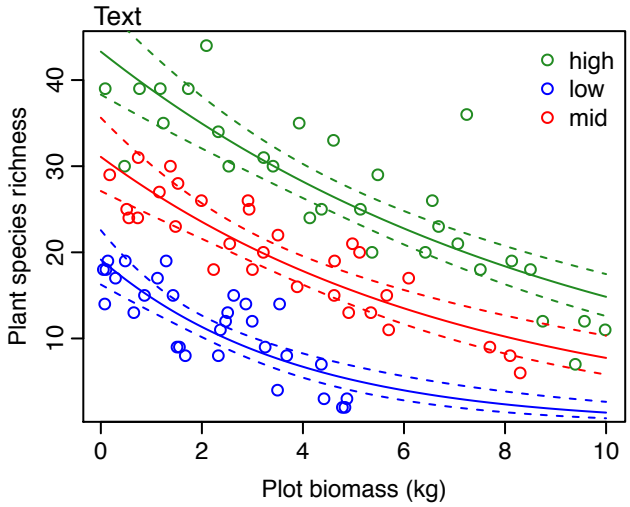



**Figure 3:** *Species richness in grassland plots as a function of biomass and soil pH. Predictions of a generalised linear model (solid line) are shown with 95% confidence limits (dashed lines) for high (green), intermediate (red) and low (blue) pH.*