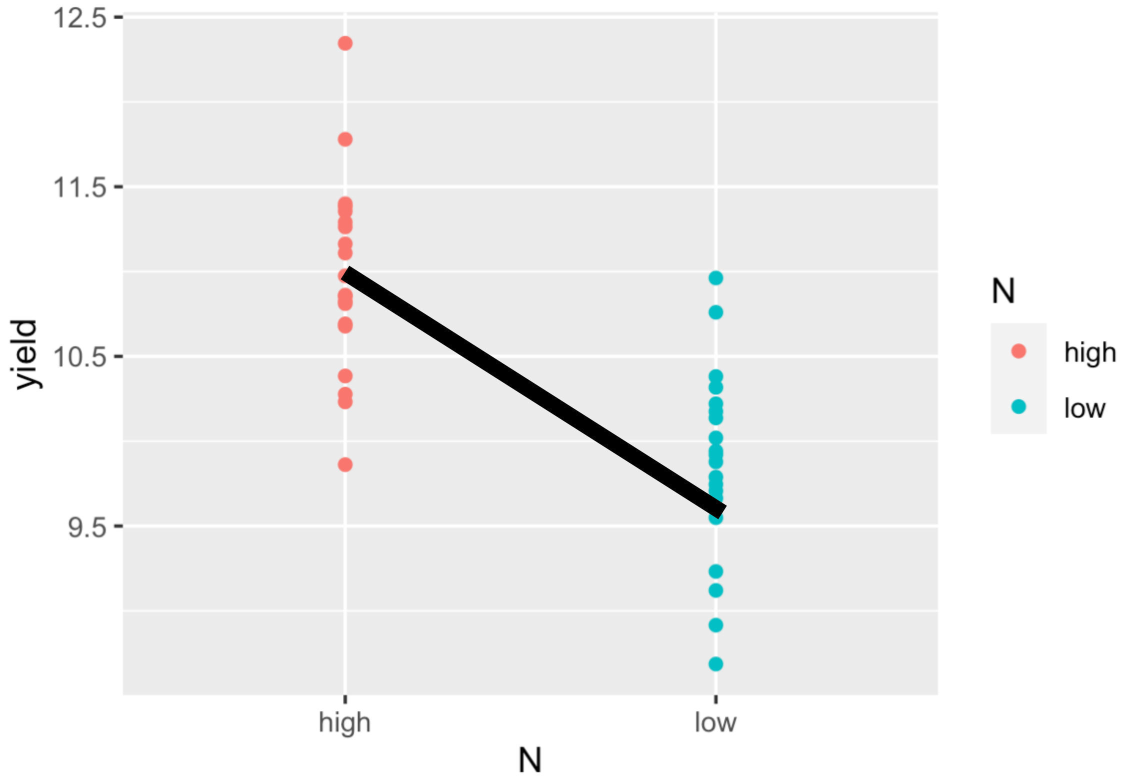
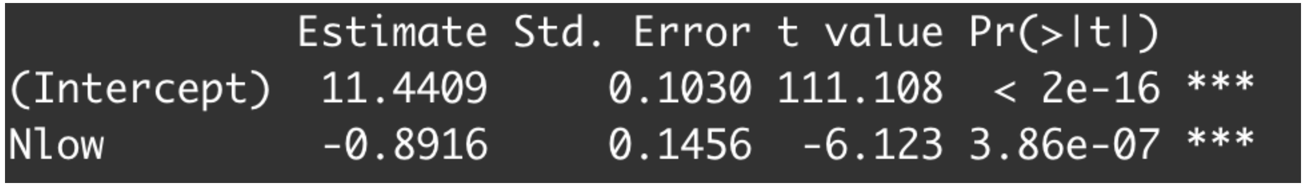
More Linear regression

- Multiple categorical variables
- Interpreting model summaries in R
- Interactions between multiple categorical variables
- Predicting categorical variable (two levels) with logistic regression
- ANOVA types

One categorical variable



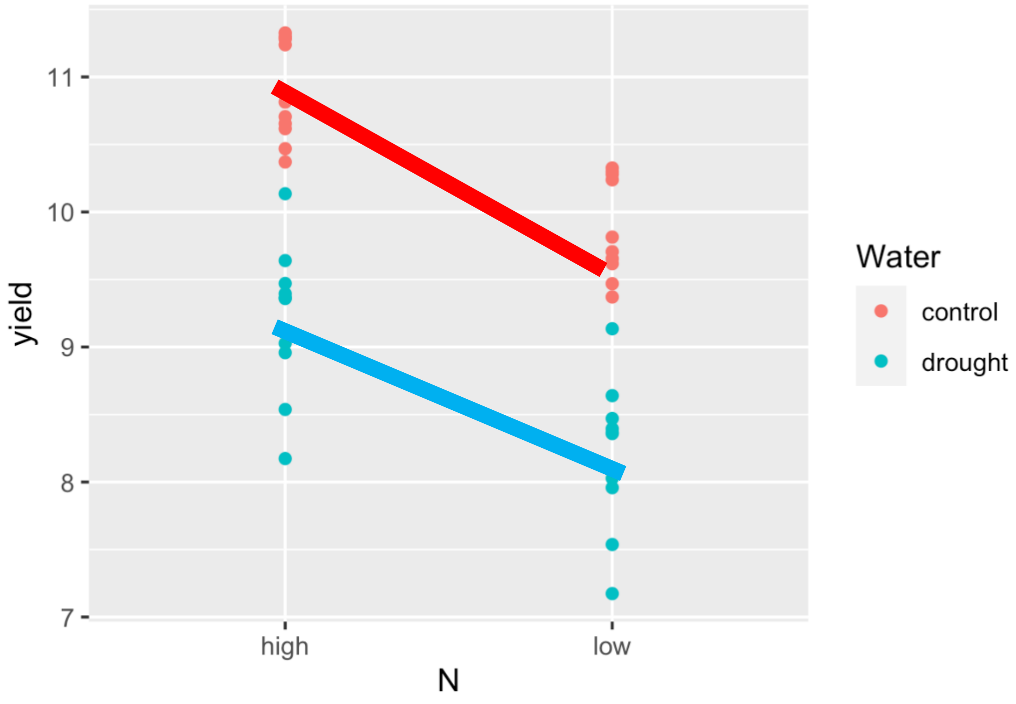


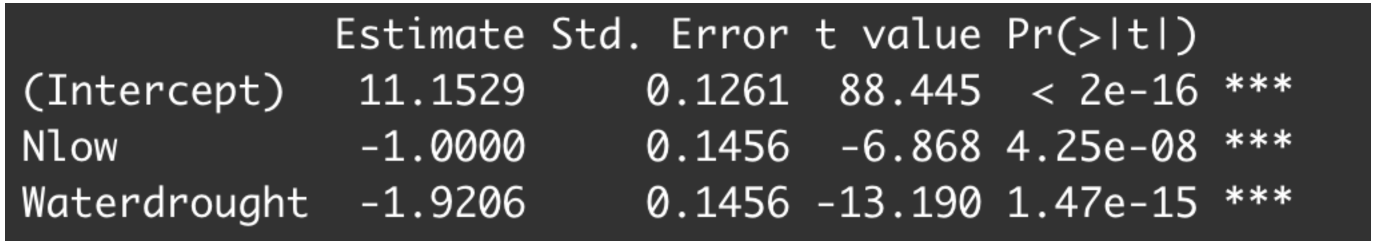
```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    11.4409     0.1030 111.108  < 2e-16 ***
Nlow           -0.8916     0.1456  -6.123 3.86e-07 ***
```

The mean of yield in high N is 11.44, which is significantly different than 0

The difference between group low and high group a is -0.89, which is significantly different than 0

Two categorical variables





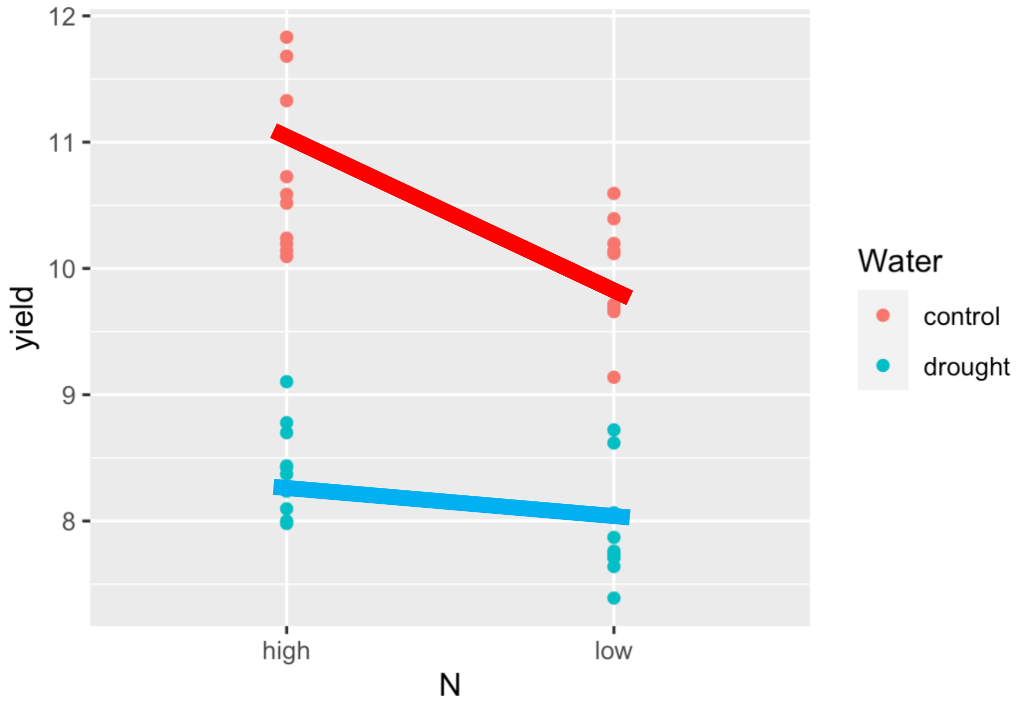
```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    11.1529     0.1261  88.445  < 2e-16 ***
Nlow           -1.0000     0.1456  -6.868 4.25e-08 ***
Waterdrought   -1.9206     0.1456 -13.190 1.47e-15 ***
```

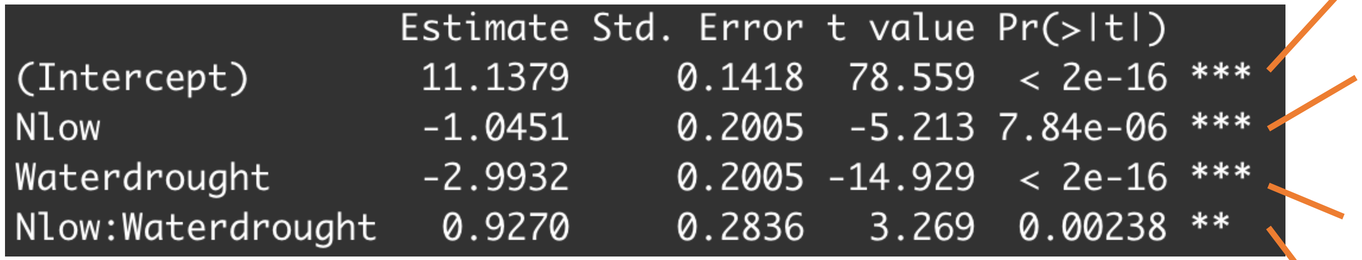
The mean yield under of high N, control

The difference between low N and high N

The difference between drought and control

Interaction between two categorical variables





```
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        11.1379     0.1418  78.559  < 2e-16 ***
Nlow               -1.0451     0.2005  -5.213 7.84e-06 ***
Waterdrought       -2.9932     0.2005 -14.929  < 2e-16 ***
Nlow:Waterdrought   0.9270     0.2836   3.269  0.00238 ** 
```

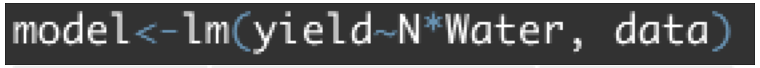
The mean yield under of high N, control

The difference between low N and high N under control

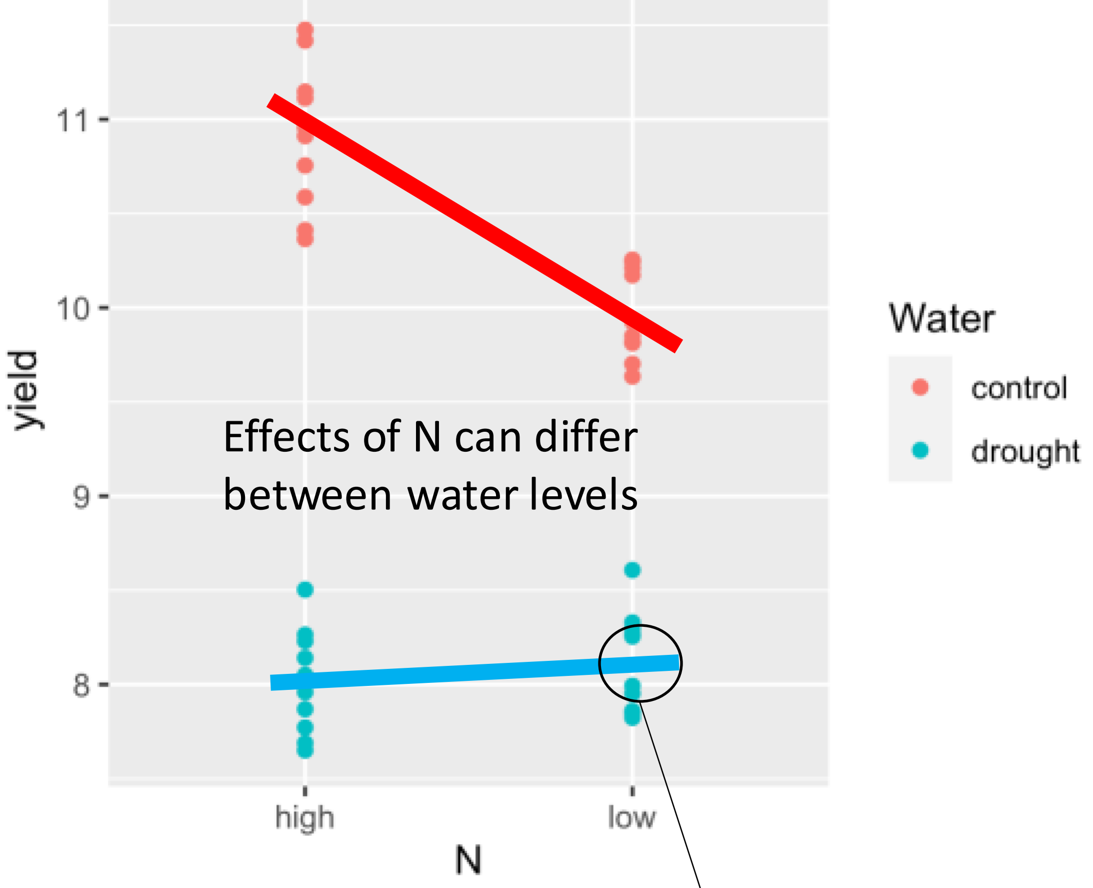
The difference between control and drought under high N

The difference in the effect of N between control and drought

With interaction term



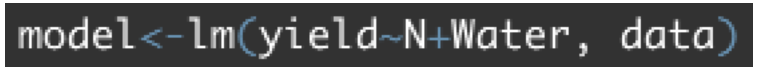
```
model<-lm(yield~N*Water, data)
```



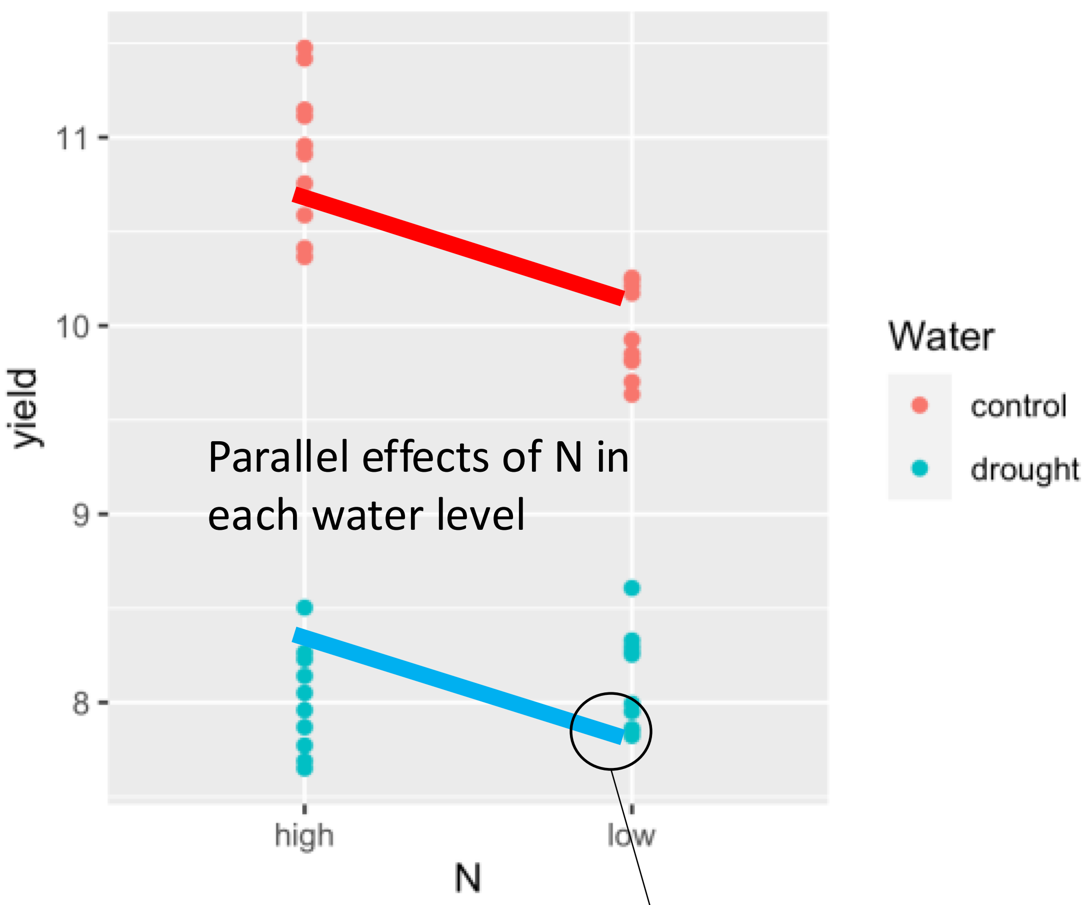
Effects of N can differ between water levels

Fitted value for yield in low N + drought
Equals mean when there is an interaction because the slope can vary

No interaction term



```
model<-lm(yield~N+Water, data)
```



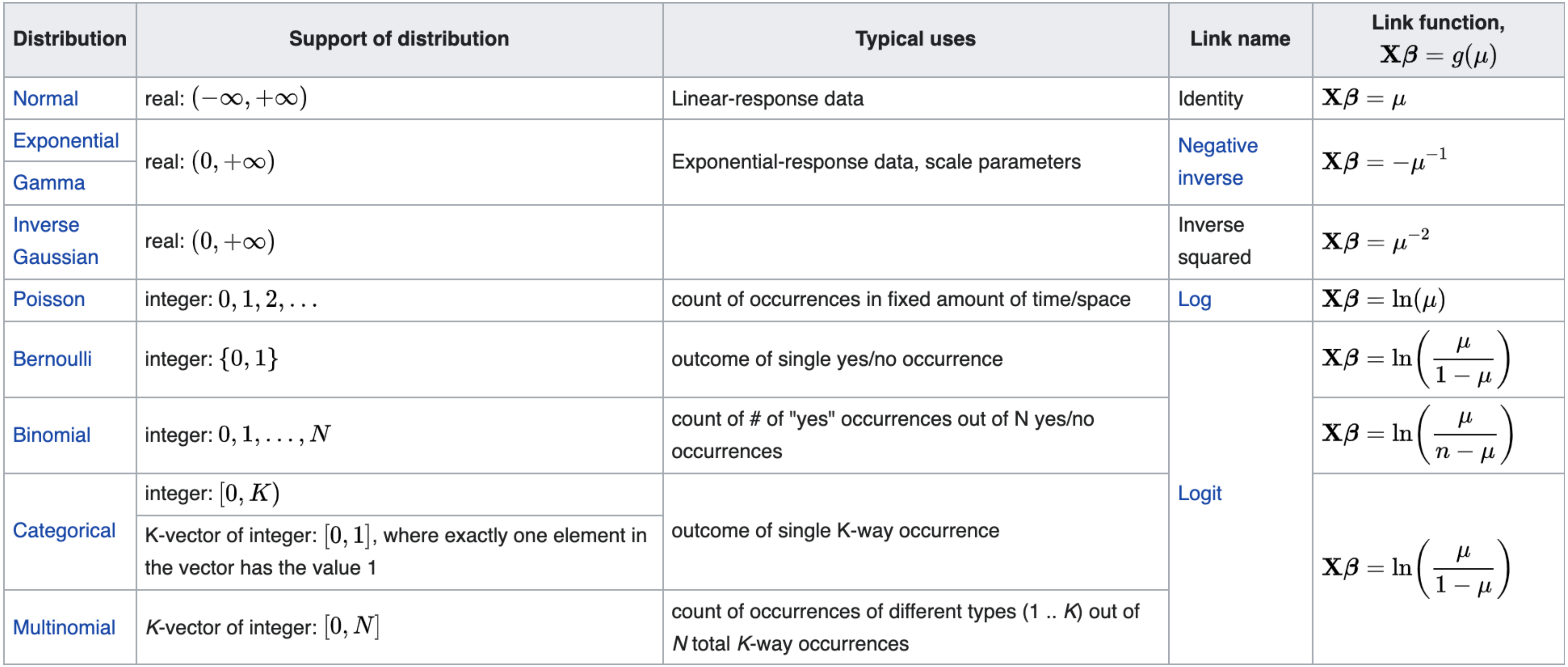
Parallel effects of N in each water level

Fitted value for yield in low N + drought
Doesn’t necessarily = mean

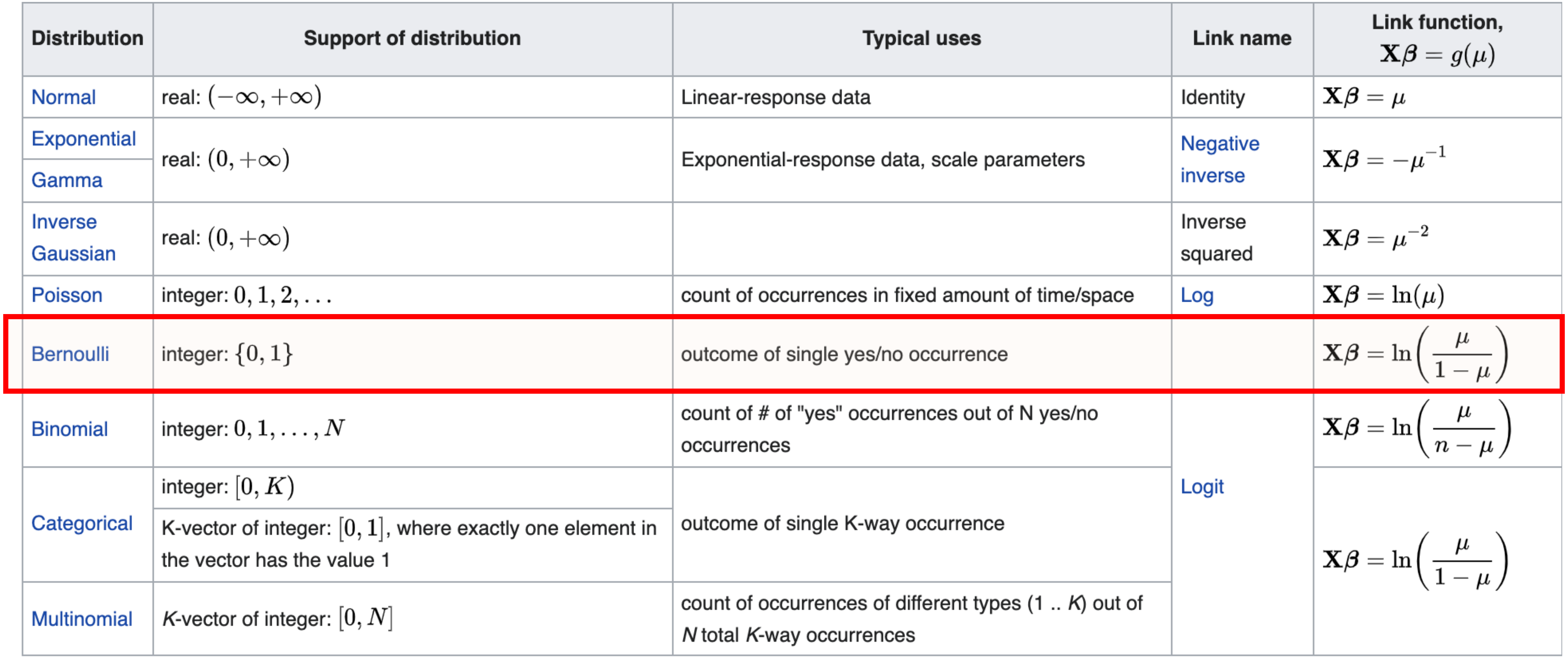
Predicting categorical variable: “Logistic regression”

(Generalized Linear Model)

What if your response variable is a category?
What if your response variable are counts?

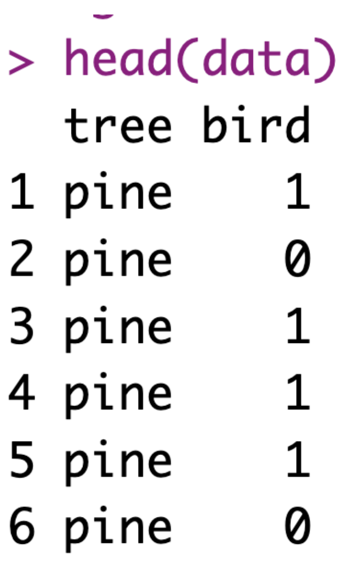


Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta}=g(\mu)$
Normal	real: $(-\infty,+\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta}=\mu$
Exponential	real: $(0,+\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta}=-\mu^{-1}$
Gamma				
Inverse Gaussian	real: $(0,+\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta}=\mu^{-2}$
Poisson	integer: $0,1,2,\ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta}=\ln(\mu)$
Bernoulli	integer: $\{0,1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta}=\ln\left(\frac{\mu}{1-\mu}\right)$
Binomial	integer: $0,1,\ldots,N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta}=\ln\left(\frac{\mu}{n-\mu}\right)$
Categorical	integer: $[0,K)$	outcome of single K-way occurrence		$\mathbf{X}\boldsymbol{\beta}=\ln\left(\frac{\mu}{1-\mu}\right)$
	K-vector of integer: $[0,1]$, where exactly one element in the vector has the value 1			
Multinomial	*K*-vector of integer: $[0,N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences		



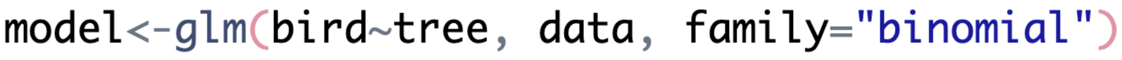
Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$
Exponential Gamma	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{n-\mu}\right)$
Categorical	integer: $[0, K)$	outcome of single K-way occurrence		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1			
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences		

Bernoulli regression with categorical predictor

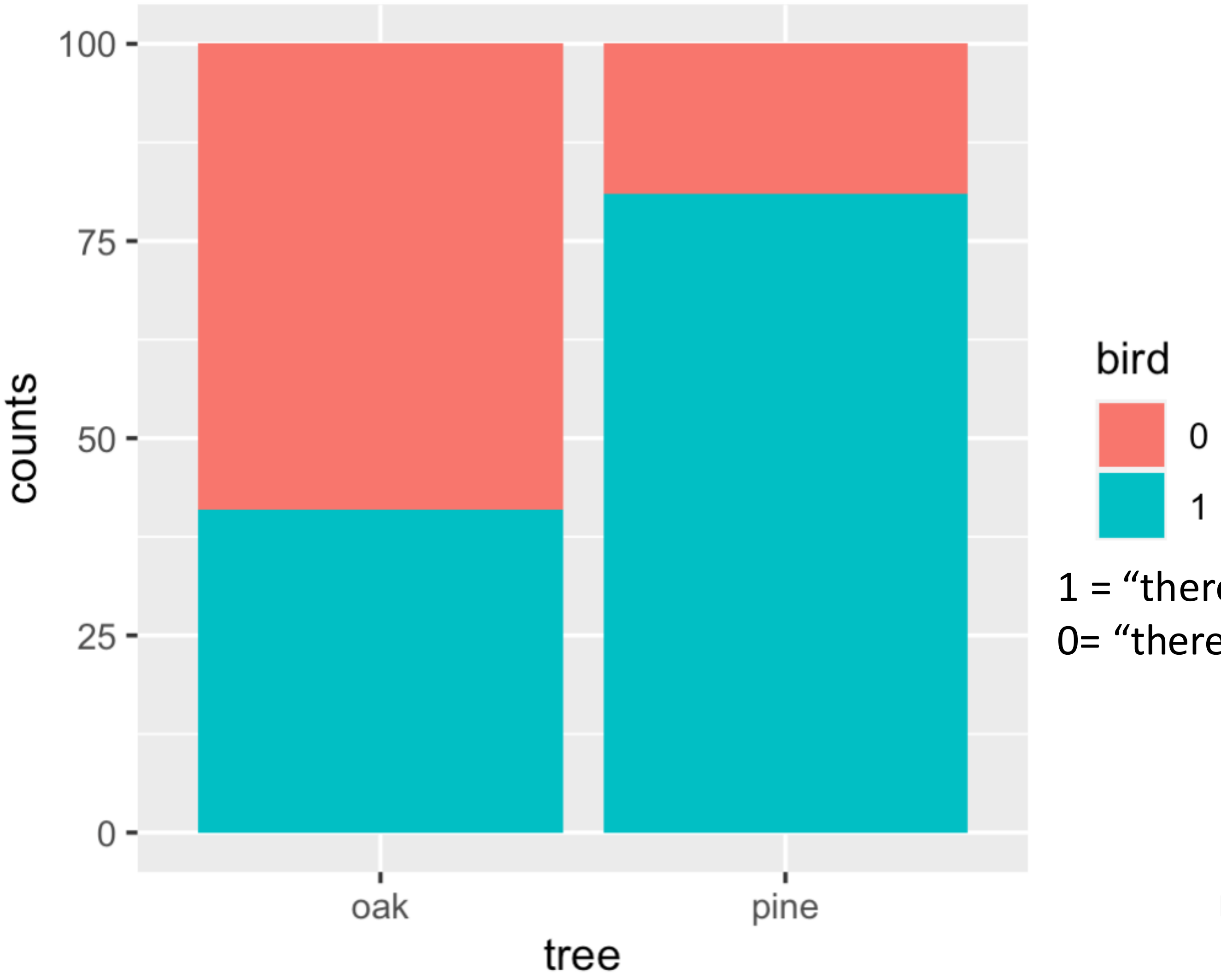


```
> head(data)
  tree bird
1 pine    1
2 pine    0
3 pine    1
4 pine    1
5 pine    1
6 pine    0
```

Are we more likely to observe a Great Salty Woodpecker in Pine trees or Oak trees?



```
model<-glm(bird~tree, data, family="binomial")
```



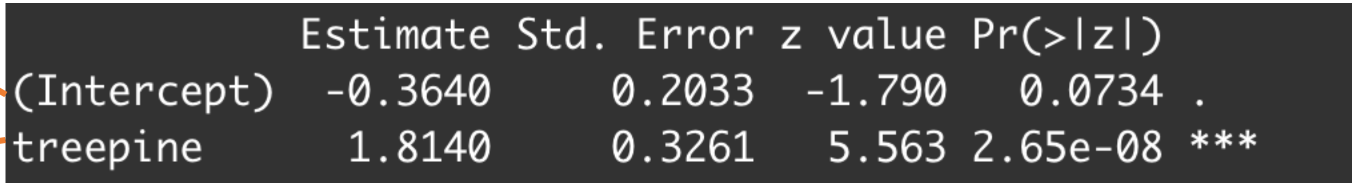
1 = “there was at least one bird”
0= “there were no birds”

Are we more likely to observe a Great Salty Woodpecker in Pine trees or Oak trees?

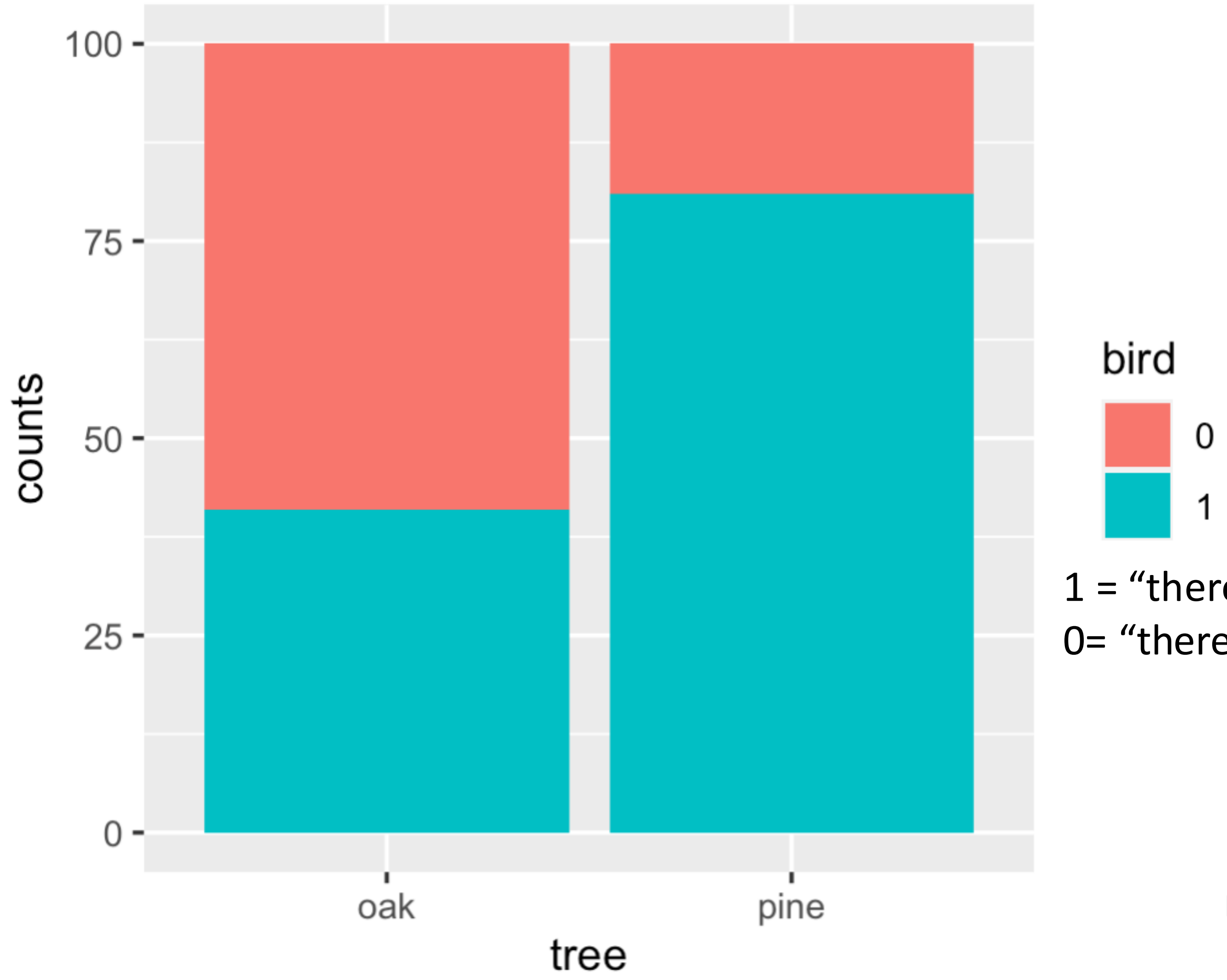
```
model<-glm(bird~tree, data, family="binomial")
```

Log odds of a bird being observed in an oak tree

Difference in Log odds of a bird being observed in a pine tree



```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3640     0.2033  -1.790   0.0734 .
treepine      1.8140     0.3261   5.563 2.65e-08 ***
```



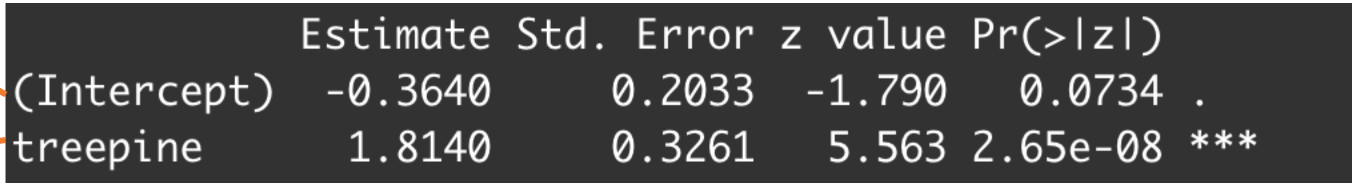
1 = “there was at least one bird”
0= “there were no birds”

Are we more likely to observe a Great Salty Woodpecker in Pine trees or Oak trees?

```
model<-glm(bird~tree, data, family="binomial")
```

Log odds of a bird being observed in an oak tree

Difference in Log odds of a bird being observed in a pine tree



```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3640     0.2033  -1.790   0.0734 .
treepine      1.8140     0.3261   5.563 2.65e-08 ***
```

Bernoulli regression with quantitative predictor

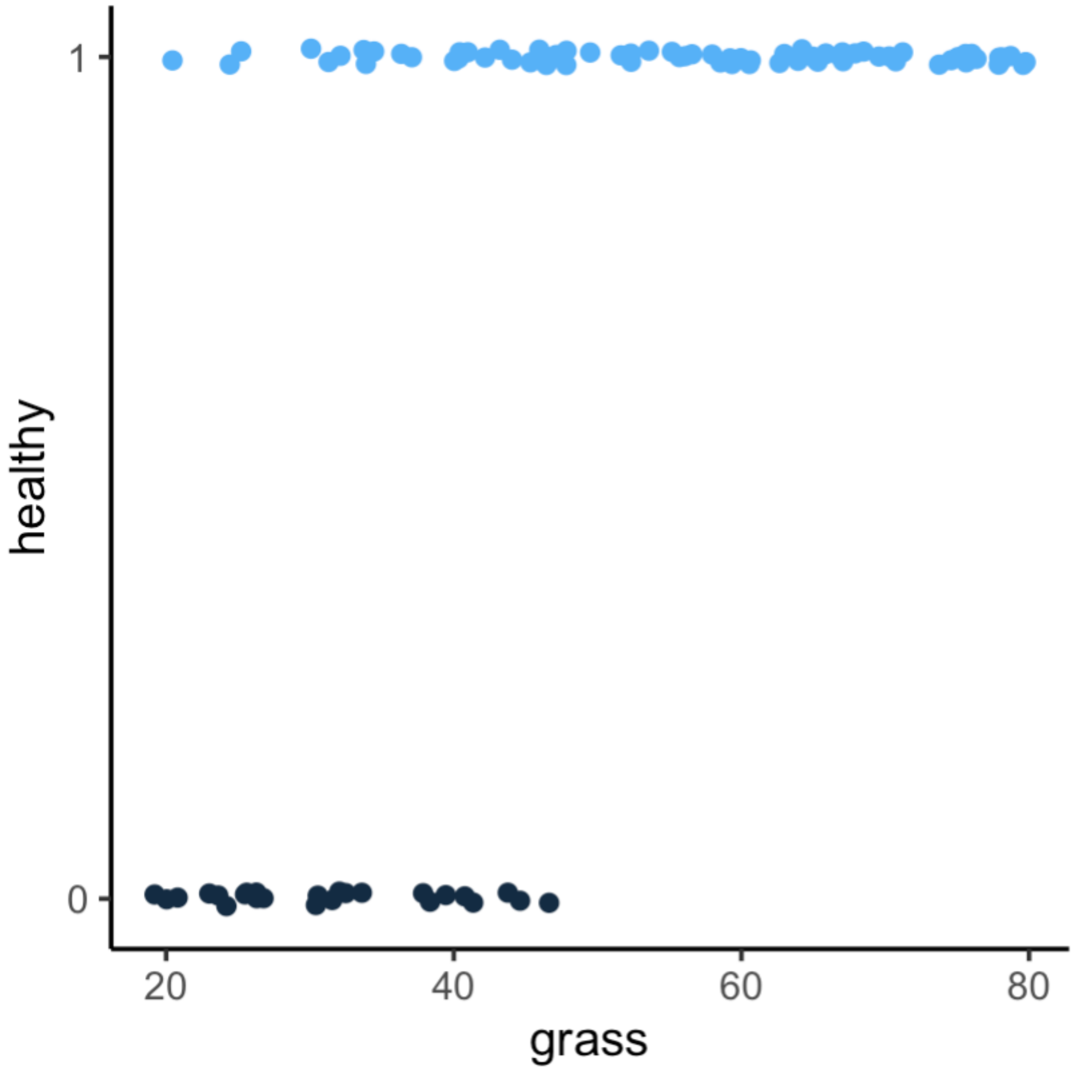
Are cows that eat more grass more likely to be scored as "healthy"?

** This is made up data!! **

Are cows that eat more grass more likely to be scored as "healthy"?

** This is made up data!! **

Grass (%)	Healthy (1=yes, 0=no)
10	1
56	0
75	1
…	…

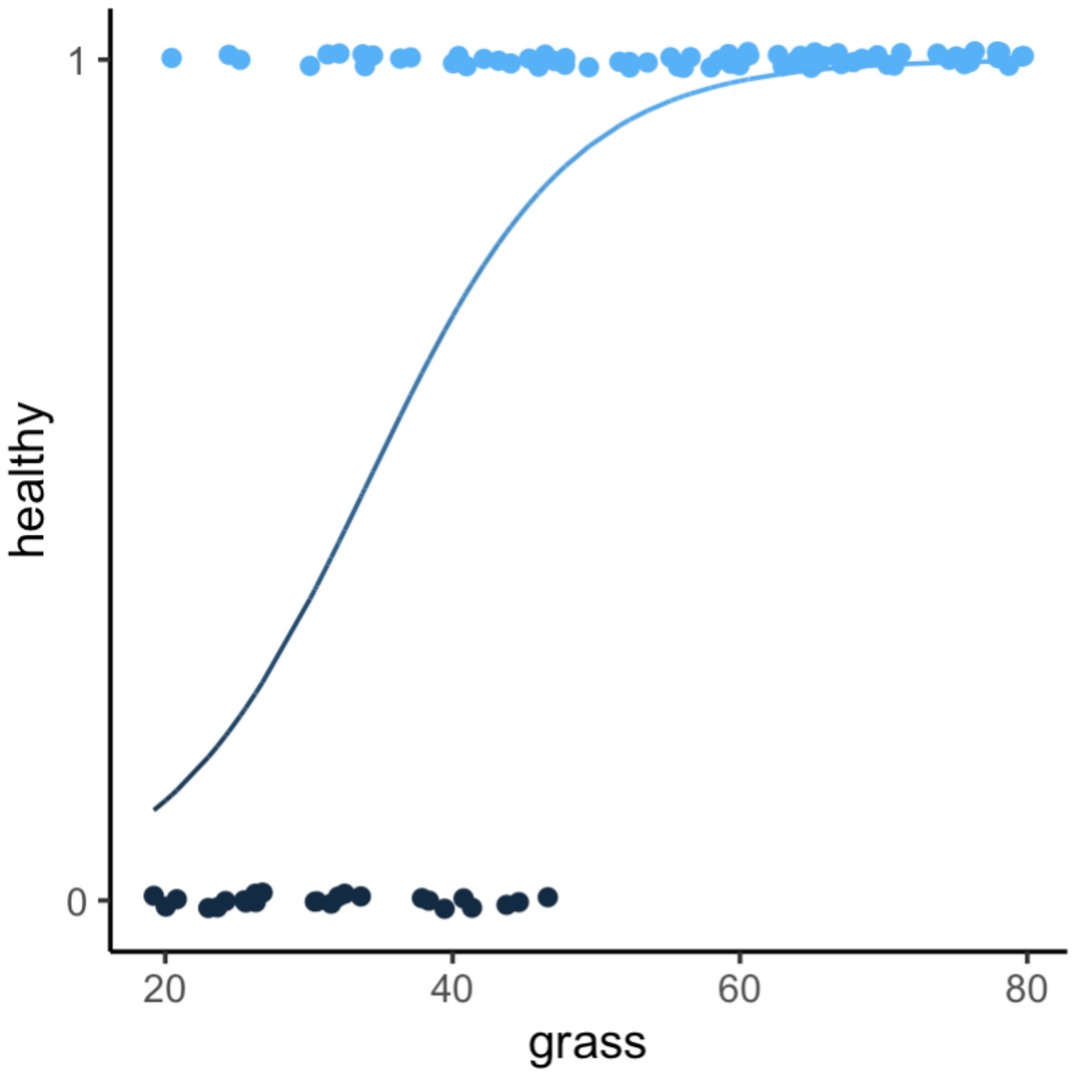


Are cows that eat more grass more likely to be scored as “healthy”?

** This is made up data!! **

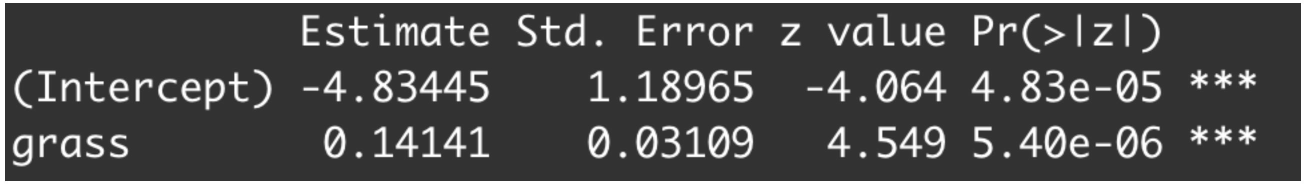
Are cows that eat more grass more likely to be scored as “healthy”?

** This is made up data!! **

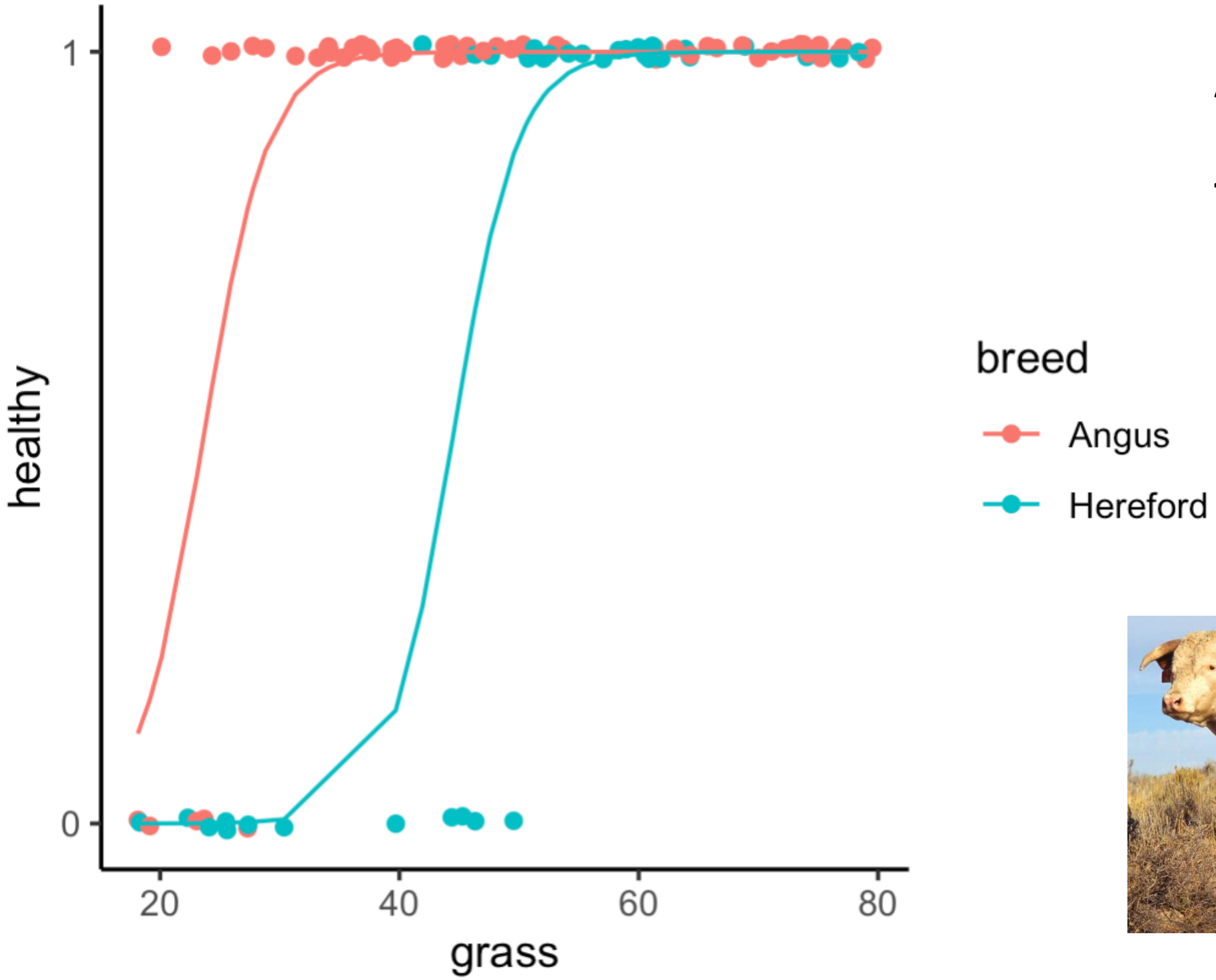


Log odds of being healthy with 0% grass diet

Increase in log odds of being healthy with each % increase in grass in diet



```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.83445    1.18965  -4.064 4.83e-05 ***
grass        0.14141    0.03109   4.549 5.40e-06 ***
```



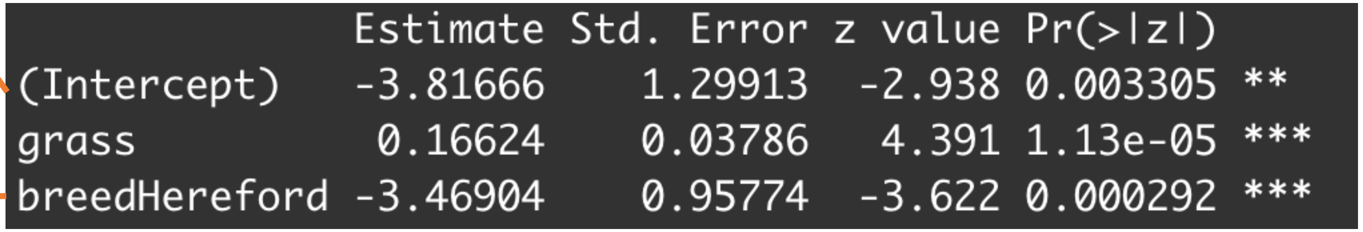
Are different breeds of cattle more likely to be scored as healthy? Even when controlling for diet?

** This is totally made up data!! **

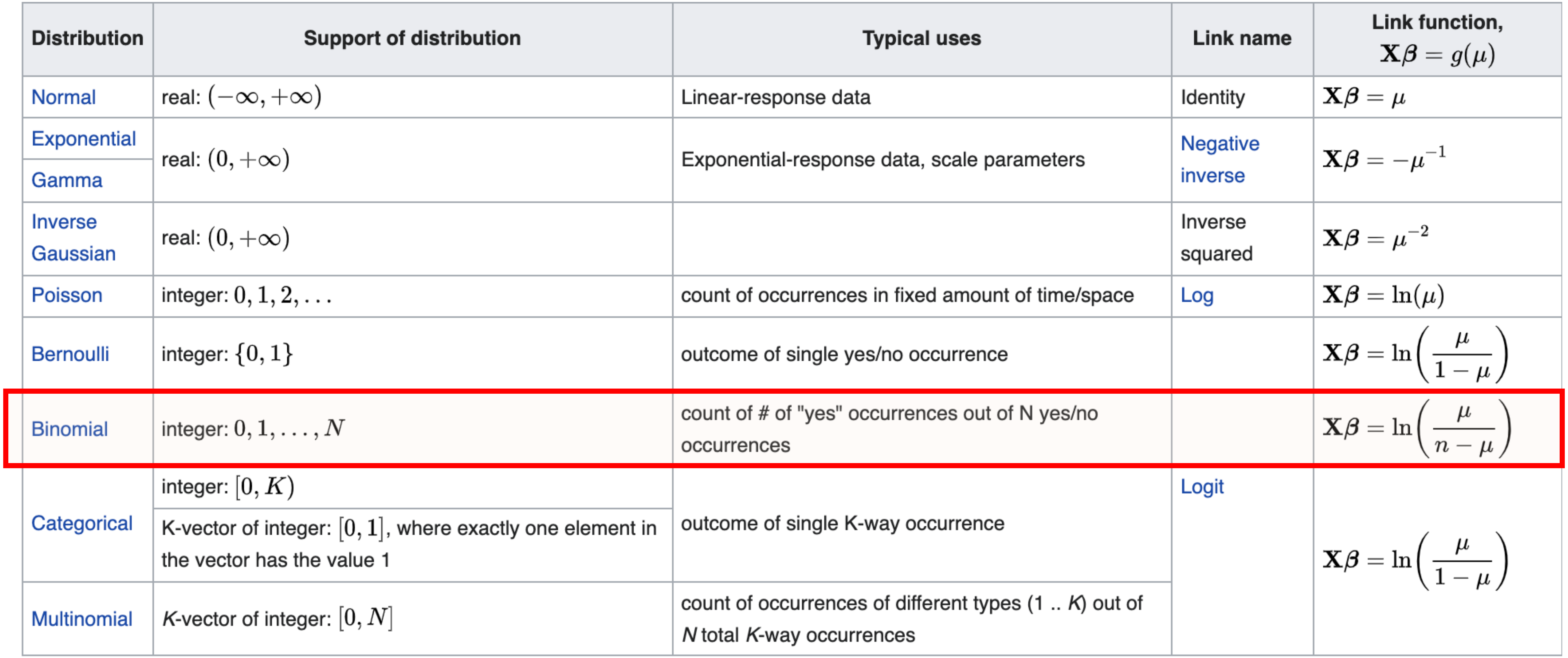
Log odds of Angus being healthy with 0% grass diet

Increase in log odds of being healthy with each % increase in grass in diet

Difference in Log odds of Hereford being healthy

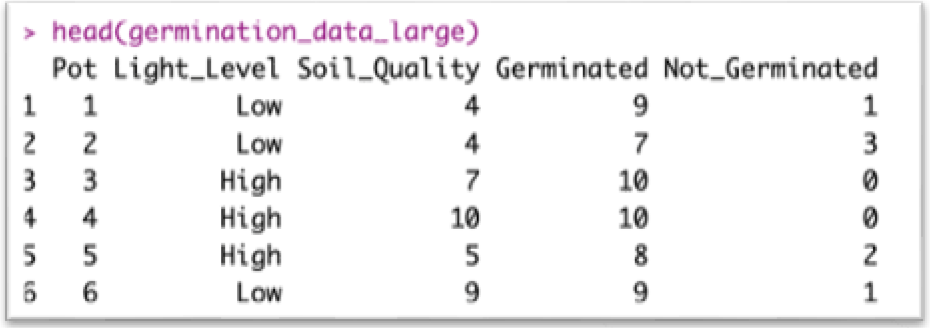


```
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -3.81666    1.29913  -2.938 0.003305 ** 
grass          0.16624    0.03786   4.391 1.13e-05 ***
breedHereford -3.46904    0.95774  -3.622 0.000292 ***
```

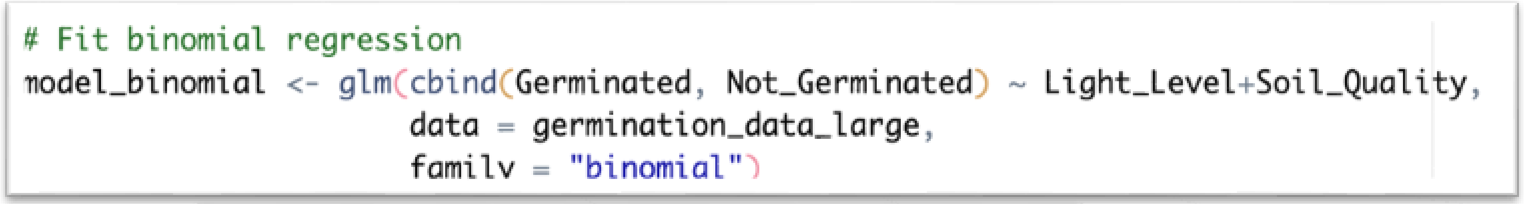


Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$
Gamma				
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{n-\mu}\right)$
Categorical	integer: $[0, K)$	outcome of single K-way occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1			
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences		

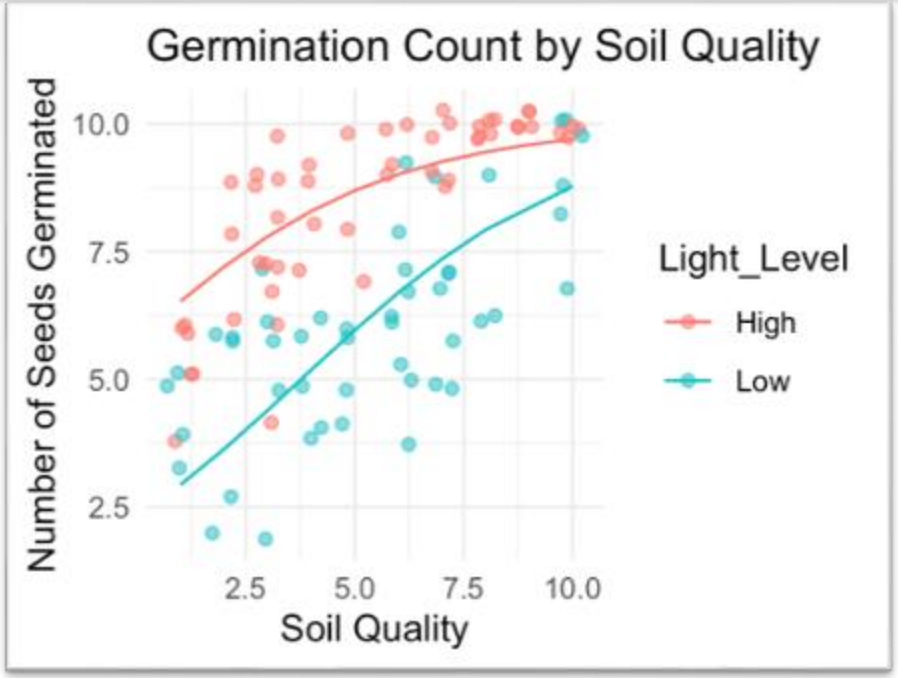
Binomial regression, from counts

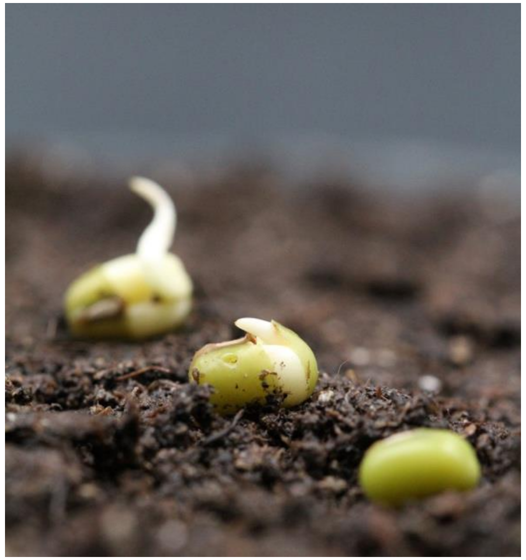


```
> head(germination_data_large)
  Pot Light_Level Soil_Quality Germinated Not_Germinated
1   1         Low            4          9              1
2   2         Low            4          7              3
3   3        High            7         10              0
4   4        High           10         10              0
5   5        High            5          8              2
6   6         Low            9          9              1
```

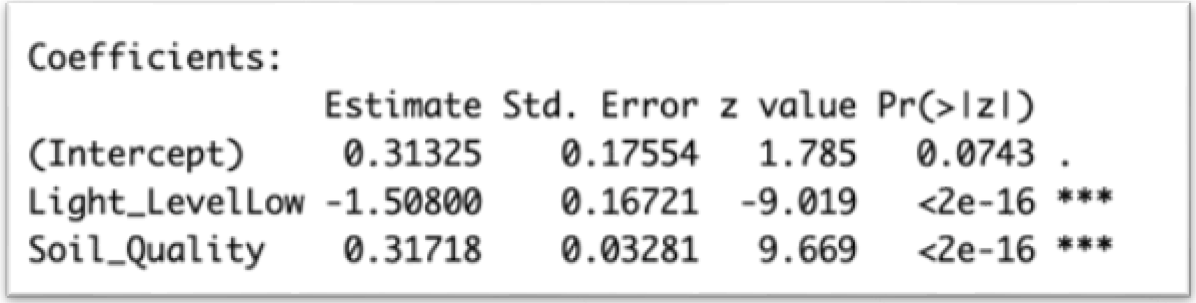


```
# Fit binomial regression
model_binomial <- glm(cbind(Germinated, Not_Germinated) ~ Light_Level+Soil_Quality,
                      data = germination_data_large,
                      family = "binomial")
```

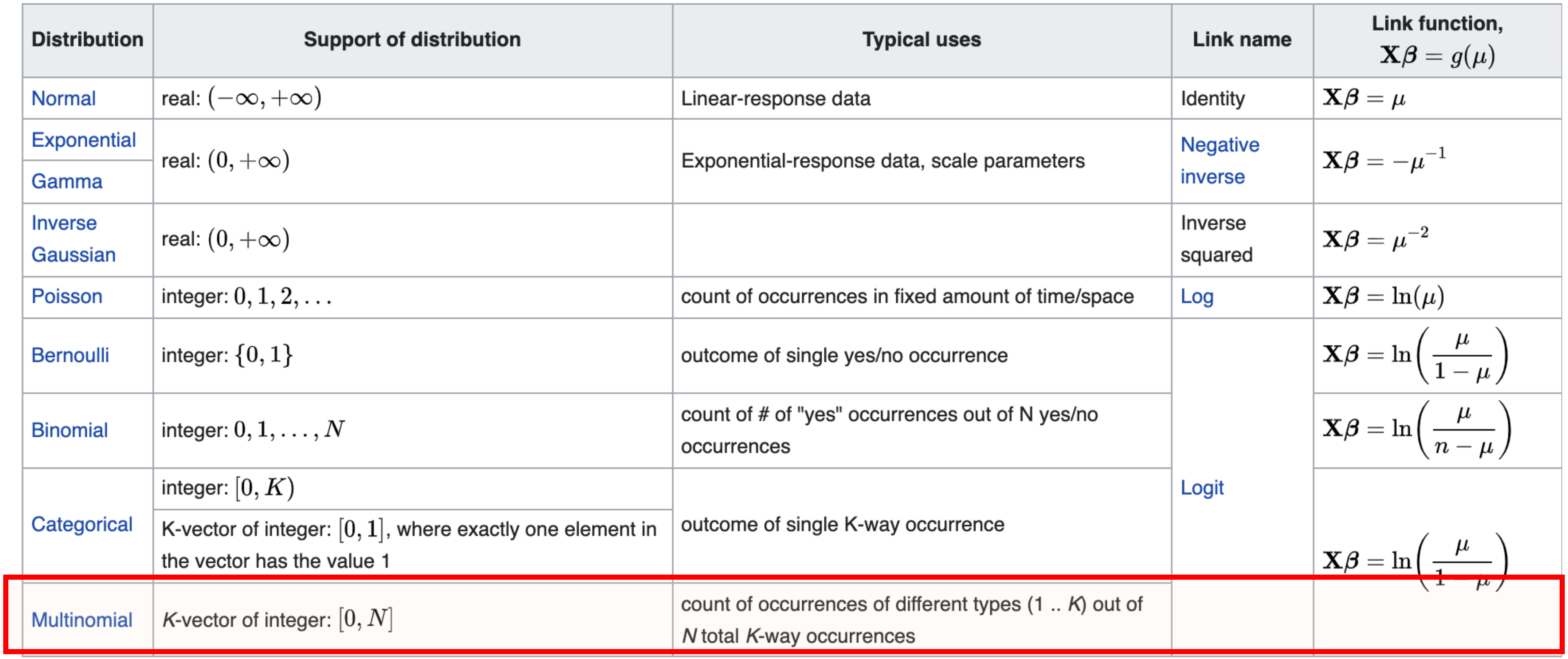




What is the effect of soil quality and light level on seed germination?



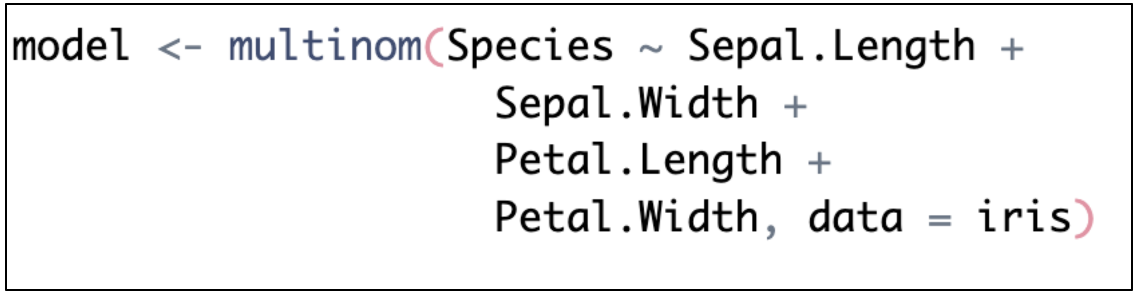
```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.31325    0.17554   1.785   0.0743 .
Light_LevelLow -1.50800    0.16721  -9.019   <2e-16 ***
Soil_Quality    0.31718    0.03281   9.669   <2e-16 ***
```



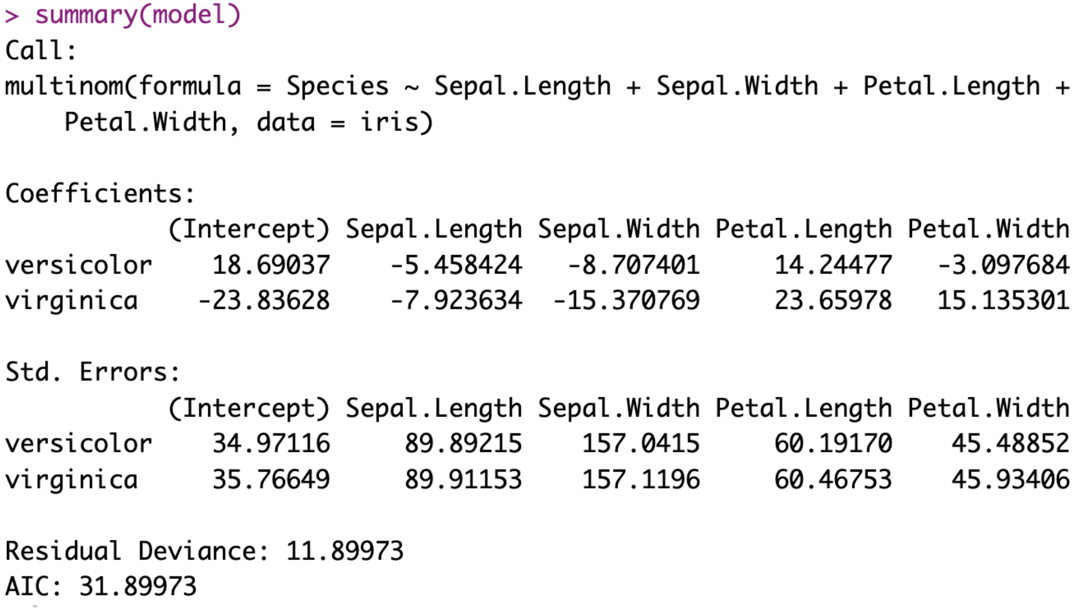
Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta}=g(\mu)$
Normal	real: $(-\infty,+\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta}=\mu$
Exponential	real: $(0,+\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta}=-\mu^{-1}$
Gamma				
Inverse Gaussian	real: $(0,+\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta}=\mu^{-2}$
Poisson	integer: $0,1,2,\ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta}=\ln(\mu)$
Bernoulli	integer: $\{0,1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta}=\ln\left(\frac{\mu}{1-\mu}\right)$
Binomial	integer: $0,1,\ldots,N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta}=\ln\left(\frac{\mu}{n-\mu}\right)$
Categorical	integer: $[0,K)$	outcome of single K-way occurrence		$\mathbf{X}\boldsymbol{\beta}=\ln\left(\frac{\mu}{1-\mu}\right)$
	K-vector of integer: $[0,1]$, where exactly one element in the vector has the value 1			
Multinomial	*K*-vector of integer: $[0,N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences		

Multinomial model:
Species (3 levels) ~ traits

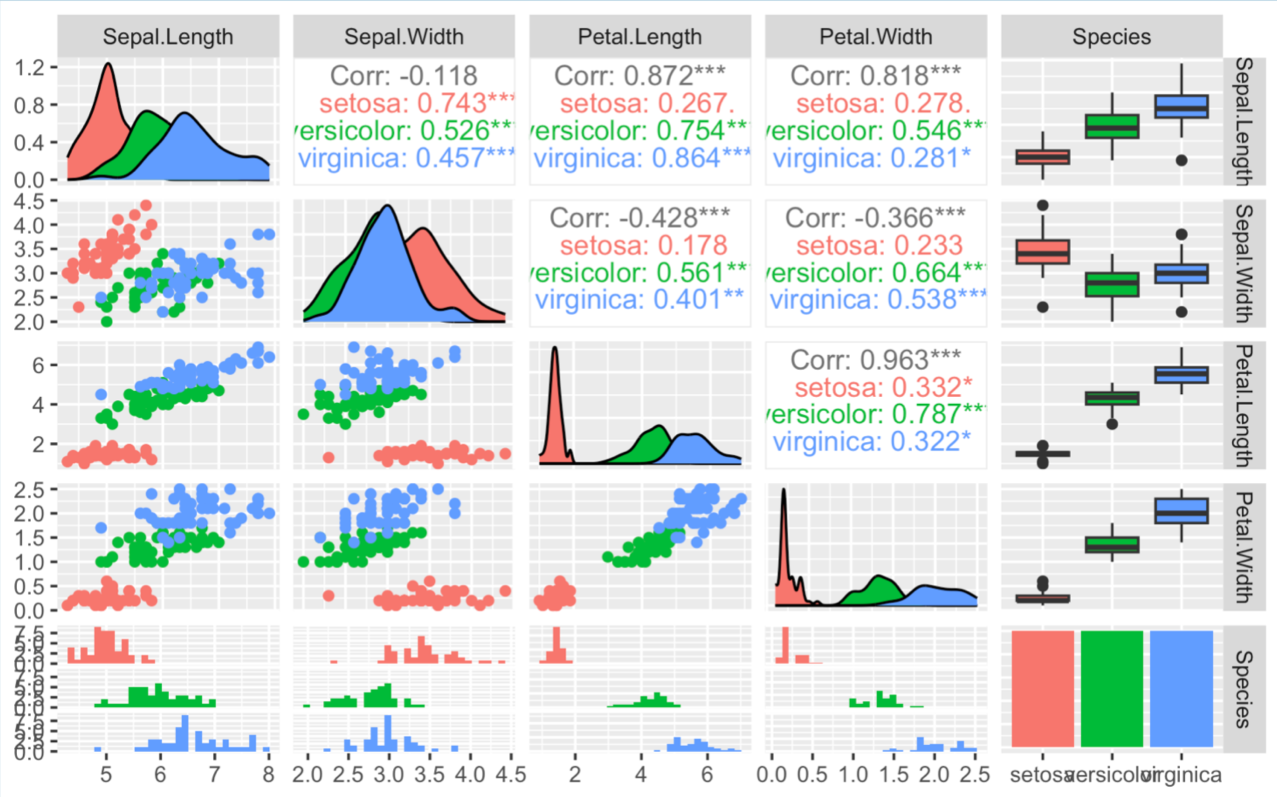
library(nnet)

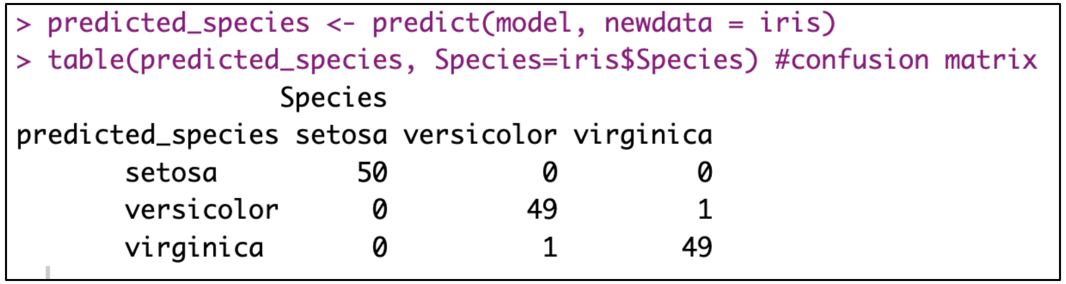


```
model <- multinom(Species ~ Sepal.Length +
                            Sepal.Width +
                            Petal.Length +
                            Petal.Width, data = iris)
```

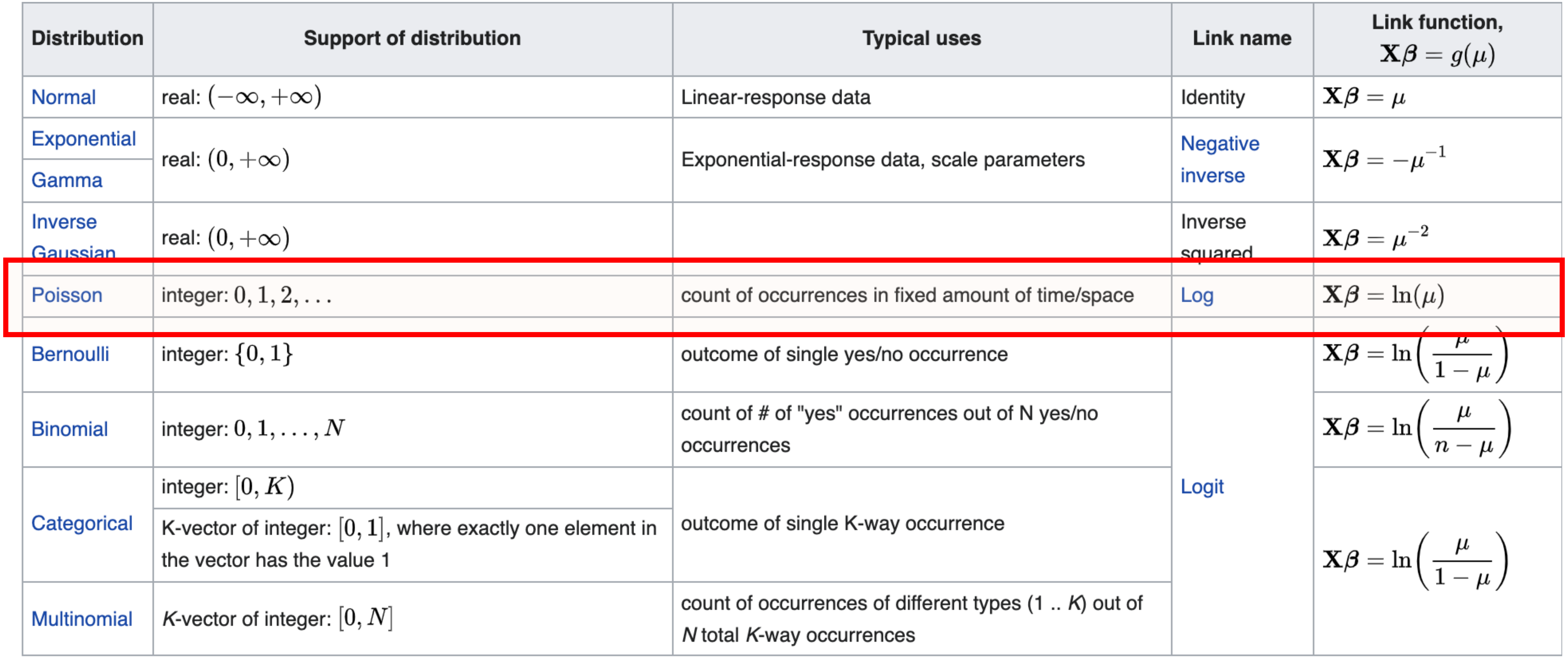


```
> summary(model)
Call:
multinom(formula = Species ~ Sepal.Length + Sepal.Width + Petal.Length +
    Petal.Width, data = iris)

Coefficients:
           (Intercept) Sepal.Length Sepal.Width Petal.Length Petal.Width
versicolor    18.69037    -5.458424   -8.707401     14.24477   -3.097684
virginica    -23.83628    -7.923634  -15.370769     23.65978   15.135301

Std. Errors:
           (Intercept) Sepal.Length Sepal.Width Petal.Length Petal.Width
versicolor    34.97116     89.89215    157.0415     60.19170    45.48852
virginica     35.76649     89.91153    157.1196     60.46753    45.93406

Residual Deviance: 11.89973
AIC: 31.89973
```



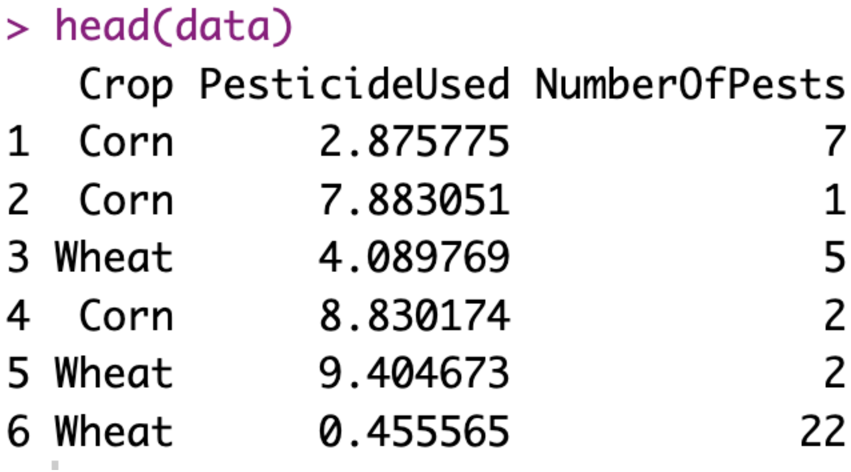


```
> predicted_species <- predict(model, newdata = iris)
> table(predicted_species, Species=iris$Species) #confusion matrix
                 Species
predicted_species setosa versicolor virginica
       setosa         50          0         0
       versicolor      0         49         1
       virginica       0          1        49
```

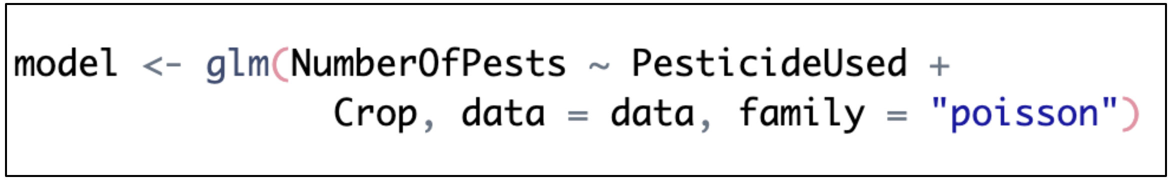


Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$
Gamma				
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{n-\mu}\right)$
Categorical	integer: $[0, K)$	outcome of single K-way occurrence		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1			
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences		

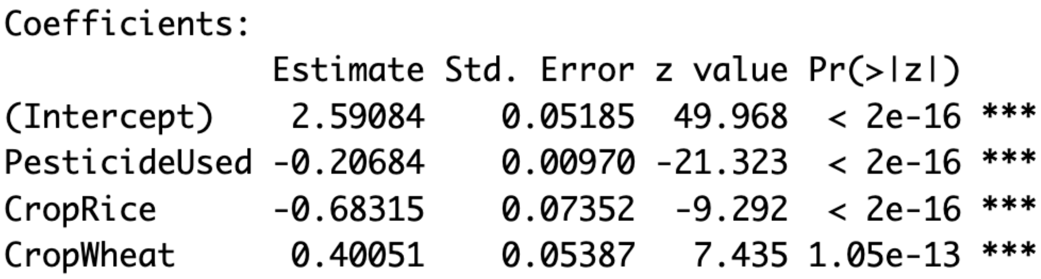
Poisson: Counts of pests ~ Pesticide + Crop



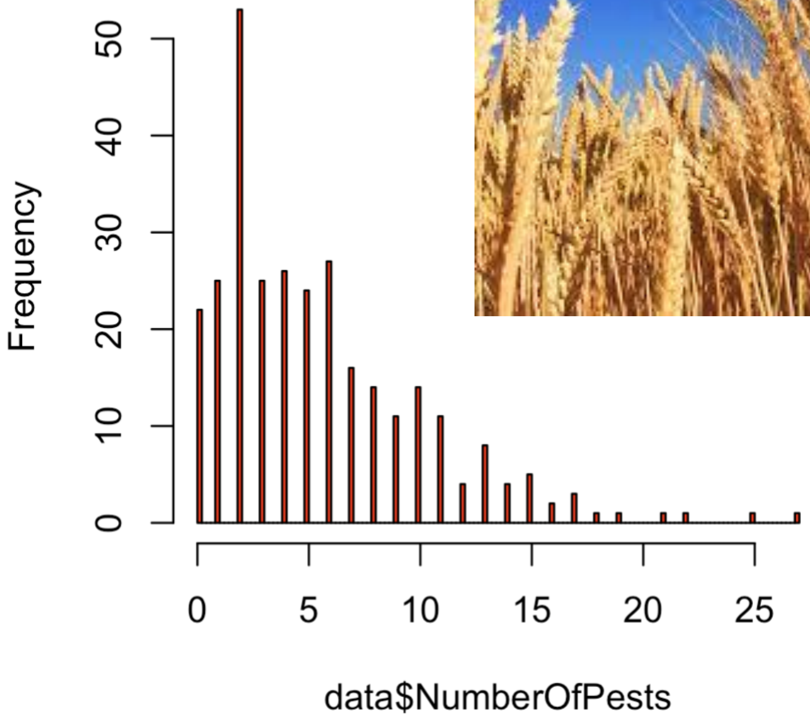
```
> head(data)
   Crop PesticideUsed NumberOfPests
1  Corn      2.875775             7
2  Corn      7.883051             1
3 Wheat      4.089769             5
4  Corn      8.830174             2
5 Wheat      9.404673             2
6 Wheat      0.455565            22
```



```
model <- glm(NumberOfPests ~ PesticideUsed +
                Crop, data = data, family = "poisson")
```



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)    2.59084    0.05185  49.968  < 2e-16 ***
PesticideUsed -0.20684    0.00970 -21.323  < 2e-16 ***
CropRice      -0.68315    0.07352  -9.292  < 2e-16 ***
CropWheat      0.40051    0.05387   7.435 1.05e-13 ***
```



ANOVA: Type I, II & III

- Type I (Sequential) SS: Use when there's a logical or theoretical sequence to the factors. For instance, if you're comparing genotypes in different fields, you might first want to consider the field effect and then the genotype effect.
- Type II SS: Use when there's no interaction between factors.
- Type III SS: Use in factorial designs, especially when interactions are of interest or when the design is unbalanced.

Type I

Grassland

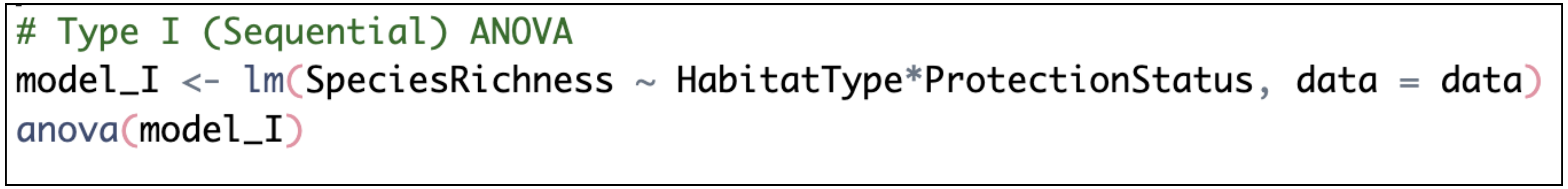
Protected

Un-Protected

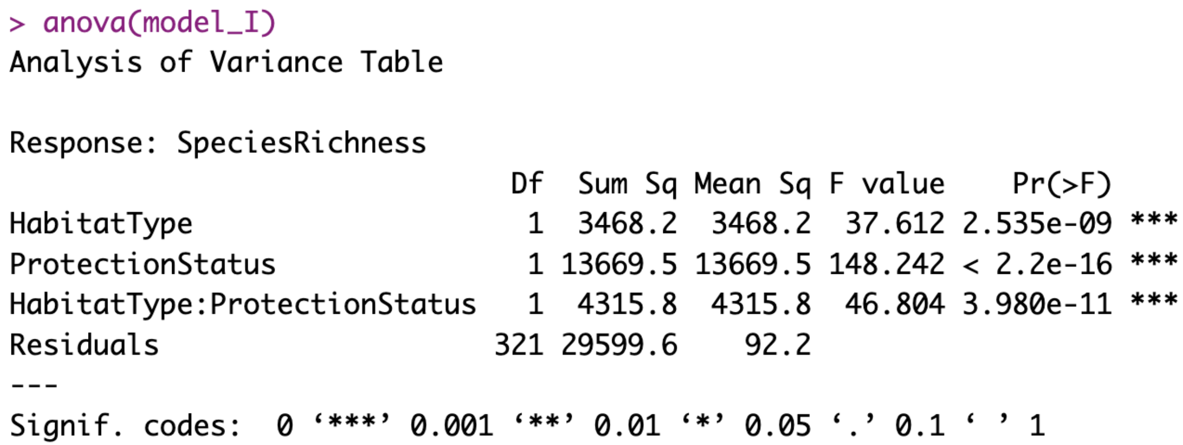
Forest

Protected

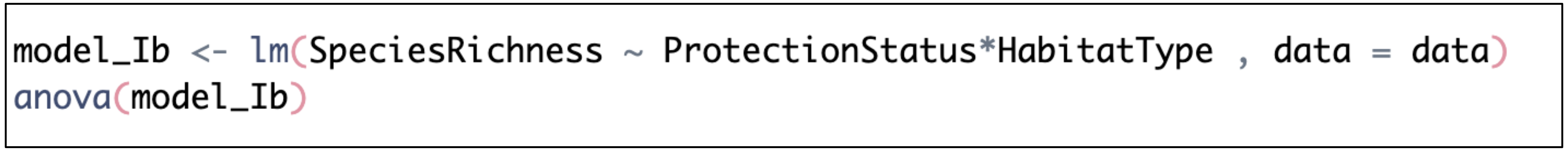
Un-Protected



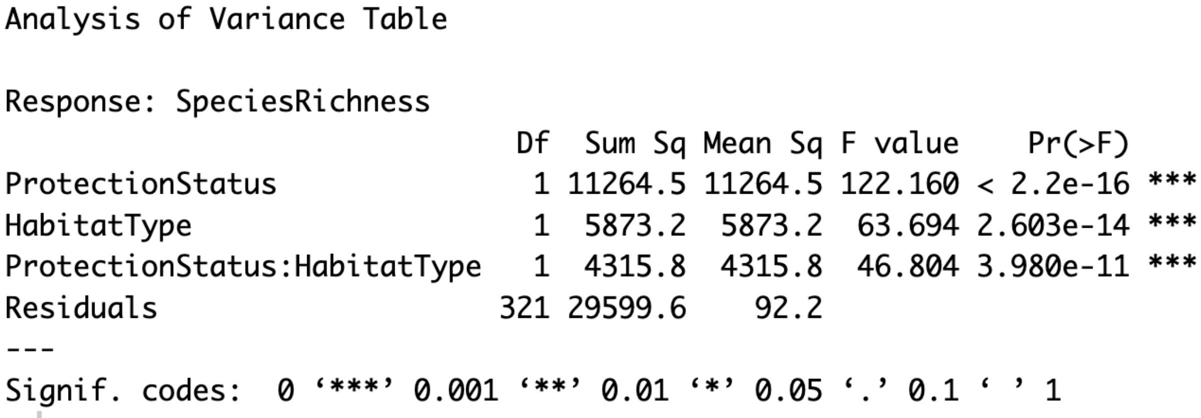
```
# Type I (Sequential) ANOVA
model_I <- lm(SpeciesRichness ~ HabitatType*ProtectionStatus, data = data)
anova(model_I)
```



```
> anova(model_I)
Analysis of Variance Table

Response: SpeciesRichness
                             Df  Sum Sq Mean Sq F value    Pr(>F)    
HabitatType                   1  3468.2  3468.2  37.612 2.535e-09 ***
ProtectionStatus              1 13669.5 13669.5 148.242 < 2.2e-16 ***
HabitatType:ProtectionStatus  1  4315.8  4315.8  46.804 3.980e-11 ***
Residuals                   321 29599.6    92.2                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
model_Ib <- lm(SpeciesRichness ~ ProtectionStatus*HabitatType , data = data)
anova(model_Ib)
```



```
Analysis of Variance Table

Response: SpeciesRichness
                             Df  Sum Sq Mean Sq F value    Pr(>F)    
ProtectionStatus              1 11264.5 11264.5 122.160 < 2.2e-16 ***
HabitatType                   1  5873.2  5873.2  63.694 2.603e-14 ***
ProtectionStatus:HabitatType  1  4315.8  4315.8  46.804 3.980e-11 ***
Residuals                   321 29599.6    92.2                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Type III

Grassland

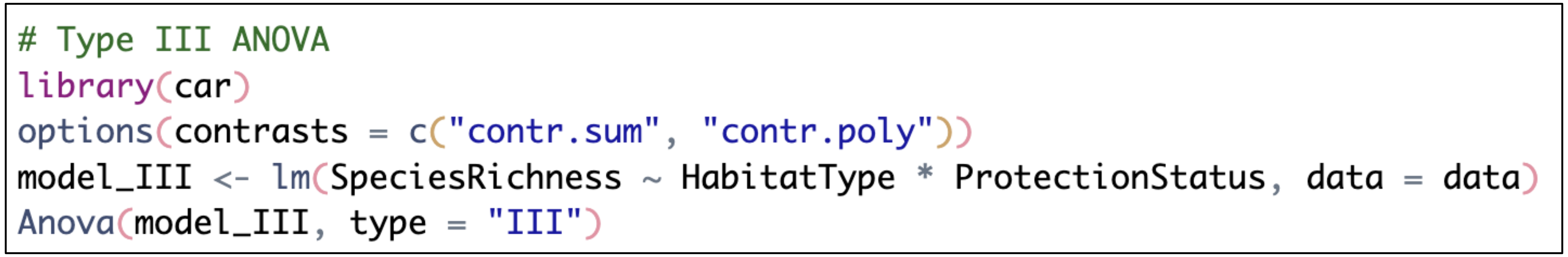
Protected

Un-Protected

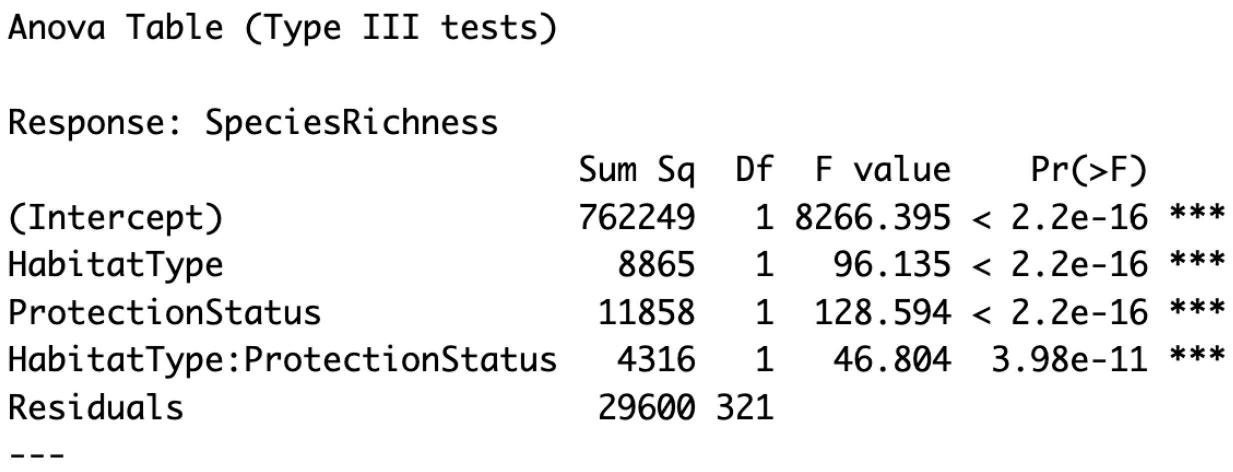
Forest

Protected

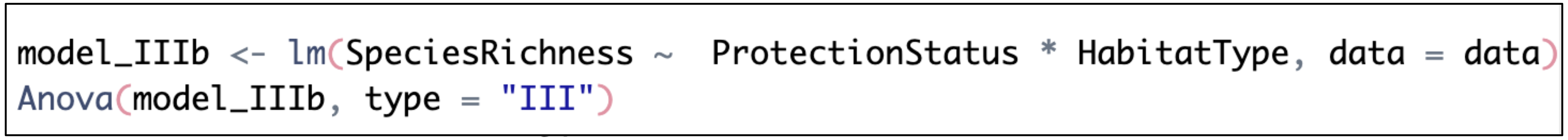
Un-Protected



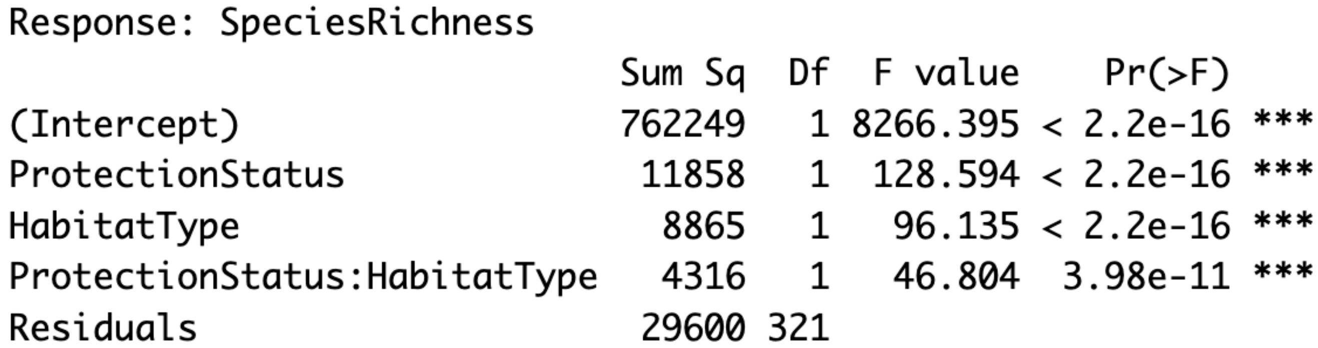
```
# Type III ANOVA
library(car)
options(contrasts = c("contr.sum", "contr.poly"))
model_III <- lm(SpeciesRichness ~ HabitatType * ProtectionStatus, data = data)
Anova(model_III, type = "III")
```



```
Anova Table (Type III tests)

Response: SpeciesRichness
                             Sum Sq  Df  F value    Pr(>F)    
(Intercept)                  762249   1 8266.395 < 2.2e-16 ***
HabitatType                    8865   1   96.135 < 2.2e-16 ***
ProtectionStatus              11858   1  128.594 < 2.2e-16 ***
HabitatType:ProtectionStatus   4316   1   46.804  3.98e-11 ***
Residuals                     29600 321                       
---
```



```
model_IIIb <- lm(SpeciesRichness ~  ProtectionStatus * HabitatType, data = data)
Anova(model_IIIb, type = "III")
```



```
Response: SpeciesRichness
                             Sum Sq  Df  F value    Pr(>F)    
(Intercept)                  762249   1 8266.395 < 2.2e-16 ***
ProtectionStatus              11858   1  128.594 < 2.2e-16 ***
HabitatType                    8865   1   96.135 < 2.2e-16 ***
ProtectionStatus:HabitatType   4316   1   46.804  3.98e-11 ***
Residuals                     29600 321                       
```