Assignment 9

Natalie Holsclaw

2025-03-23

(1) Logistic Regression

```
# Load necessary packages
library(here)
```

```
## here() starts at C:/Users/natal/OneDrive/Documents/Duke-Nat_top/Spring 2025/Statistics/Lab/nholsclaw
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(naniar)
```

```
## Warning: package 'naniar' was built under R version 4.4.3
```

```
library(DHARMa)
```

```
## Warning: package 'DHARMa' was built under R version 4.4.3
```

```
## This is DHARMa 0.4.7. For overview type '?DHARMa'. For recent changes, type news(package = 'DHARMa')
```

```
library(gtsummary)
```

```
## Warning: package 'gtsummary' was built under R version 4.4.3
```

```r
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.4.3
```

```r
# Load in data
lizard <- read_csv("jrn_lizard.csv")
```

```
## Rows: 4091 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): zone, site, plot, spp, sex, rcap, tail
## dbl  (6): pit, toe_num, SV_length, total_length, weight, pc
## date (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# filter data
sb_lizard <- lizard %>%
  filter(spp == "UTST")
```

Step 1: Define the research questions

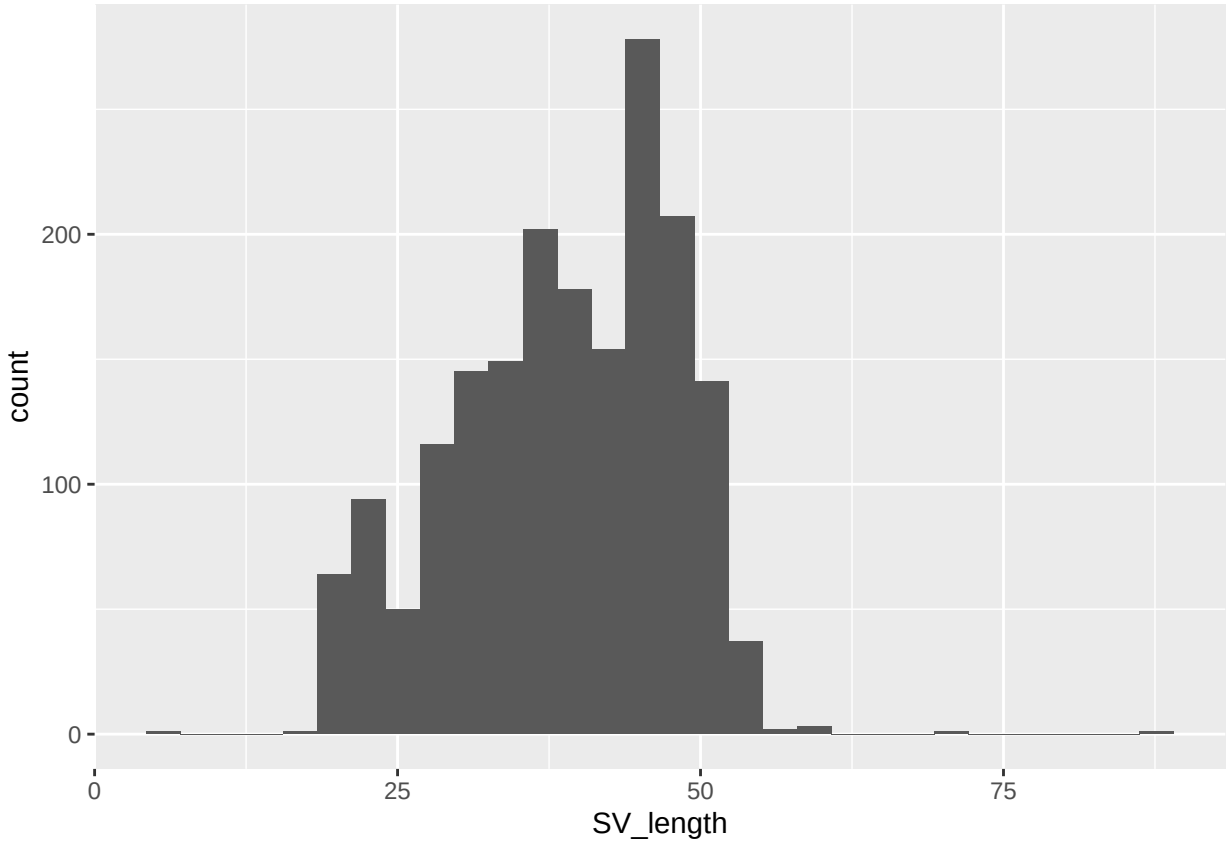
"Do snout-to-vent length, sex, and vegetation zone at time of capture significantly predict if a lizard tail is recorded as whole?"

```r
# Raw distributions
ggplot(sb_lizard, aes(x = SV_length))+
  geom_histogram()
```

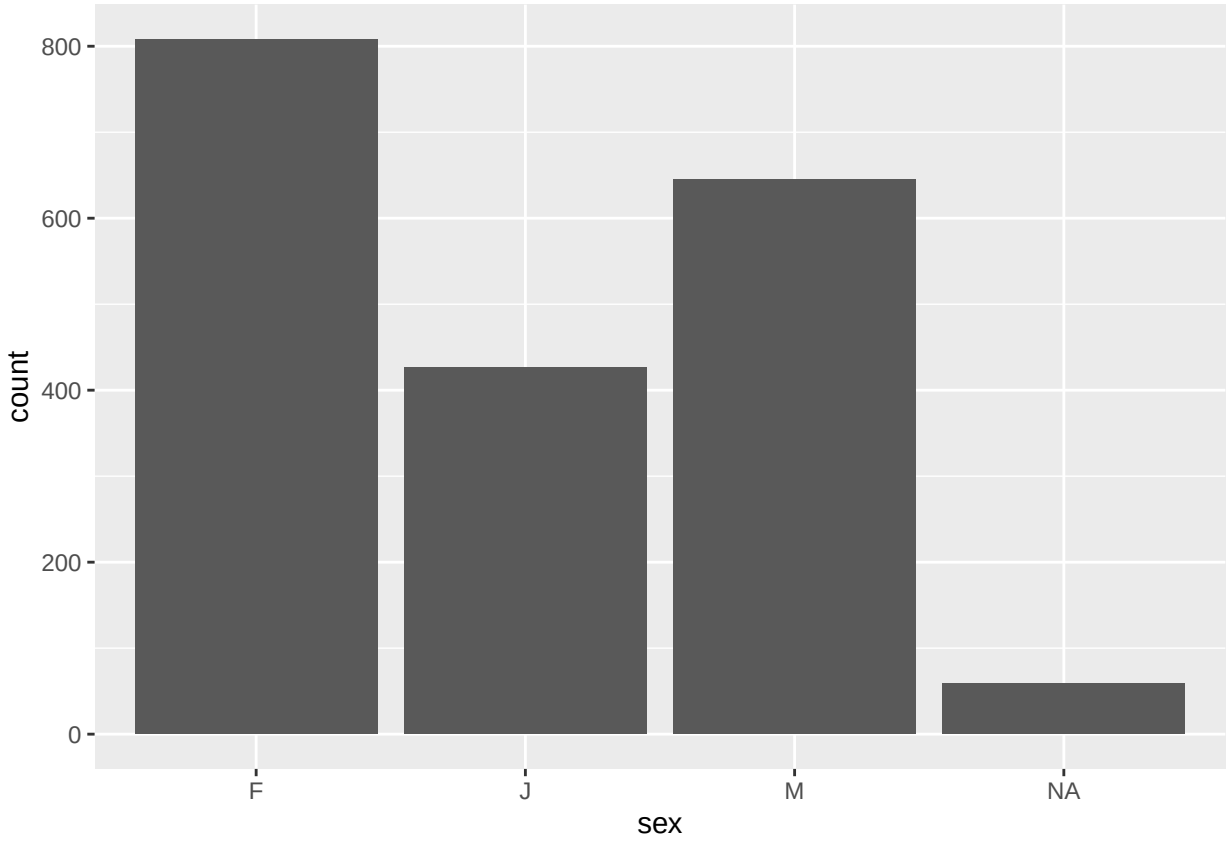
Step 2: Examine data and possible correlations

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

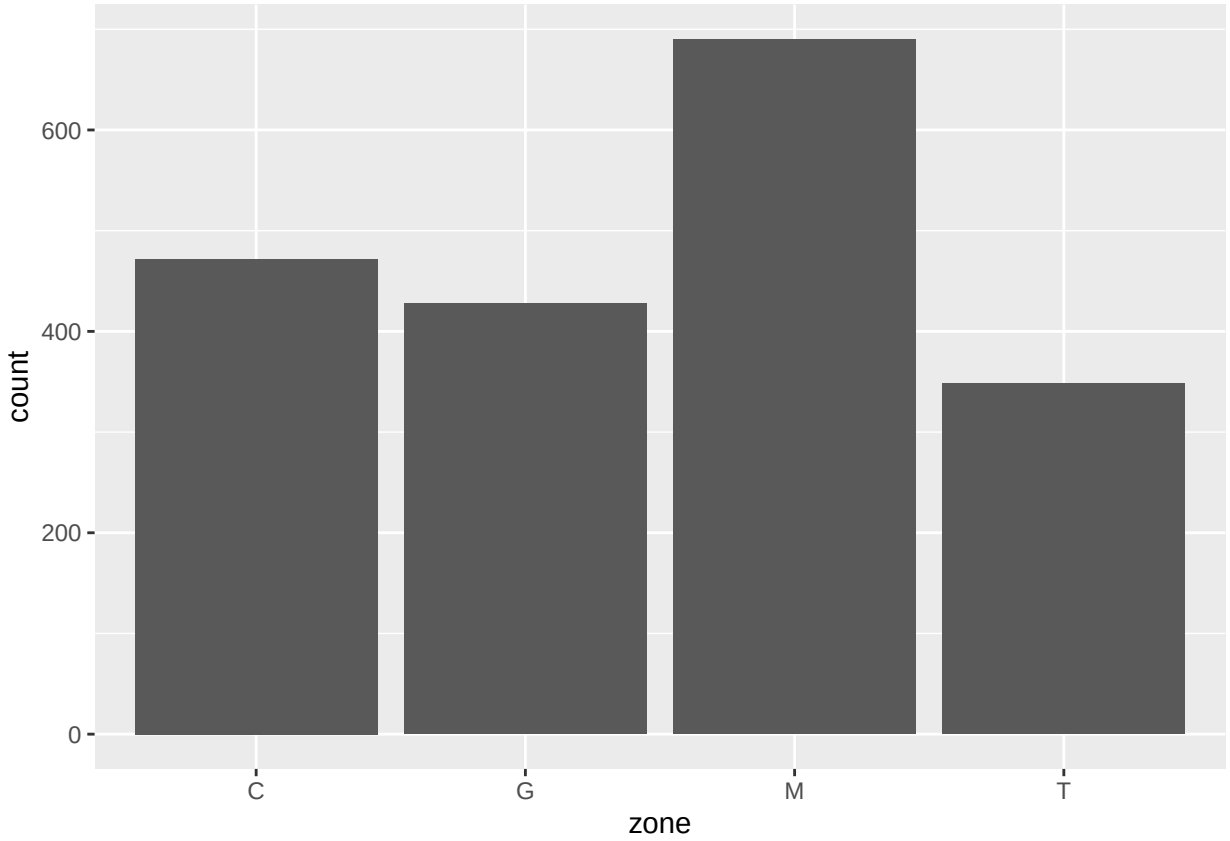
```
## Warning: Removed 114 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



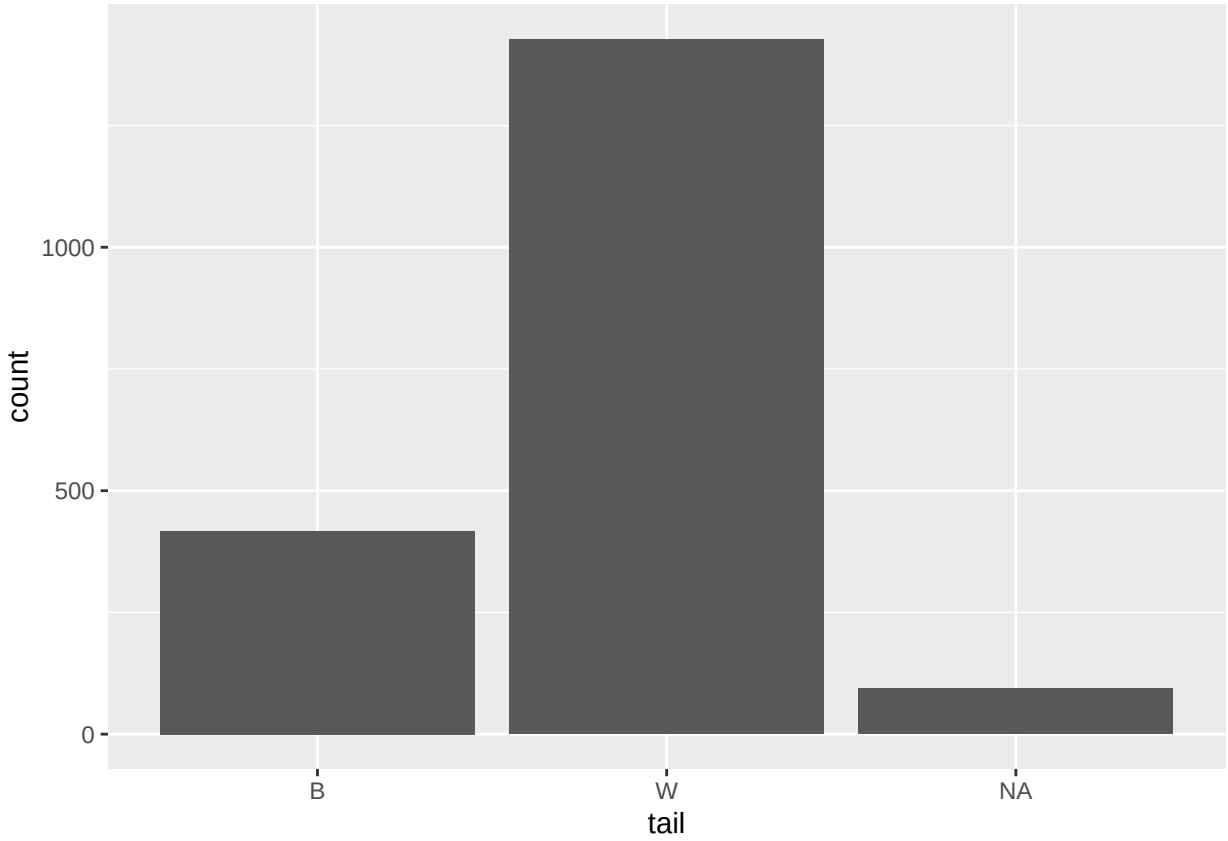

```
ggplot(sb_lizard, aes(x = sex))+
  geom_bar()
```



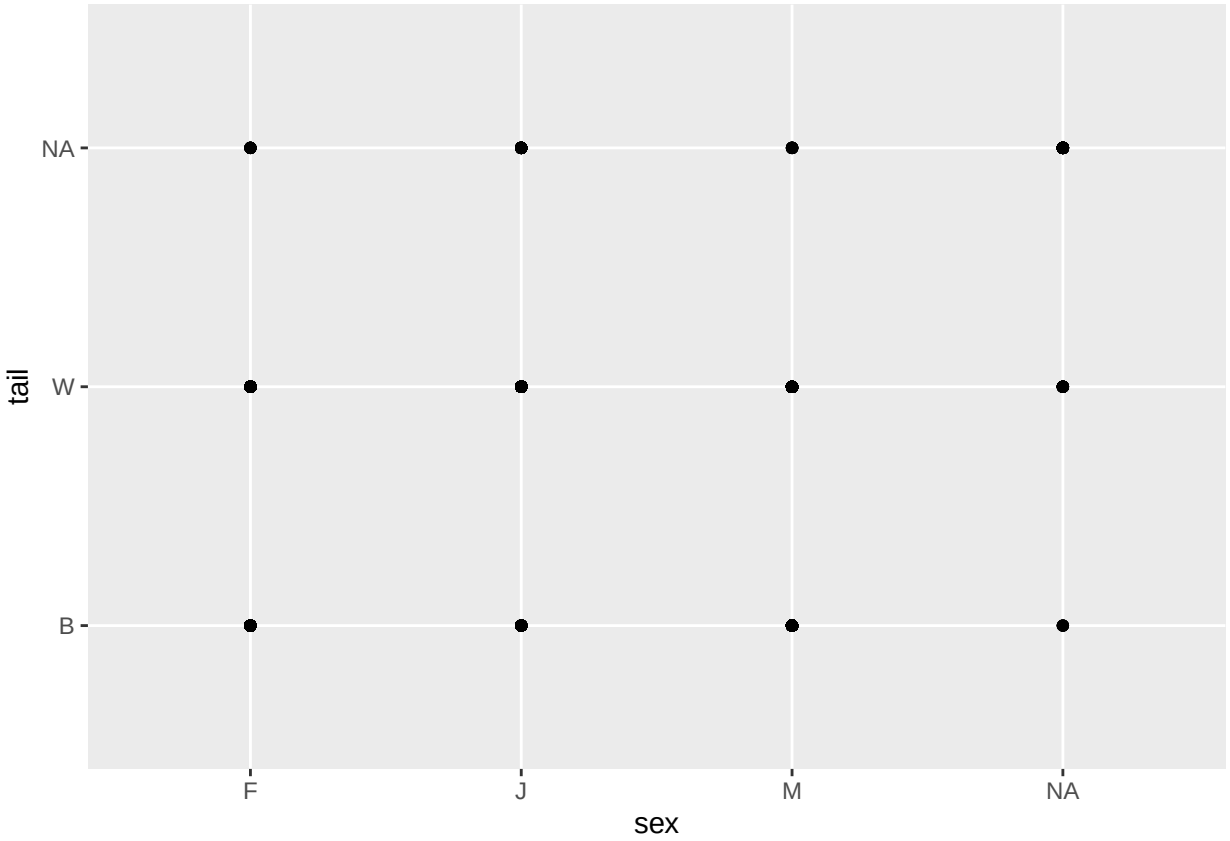

```
ggplot(sb_lizard, aes(x = zone))+
  geom_bar()
```



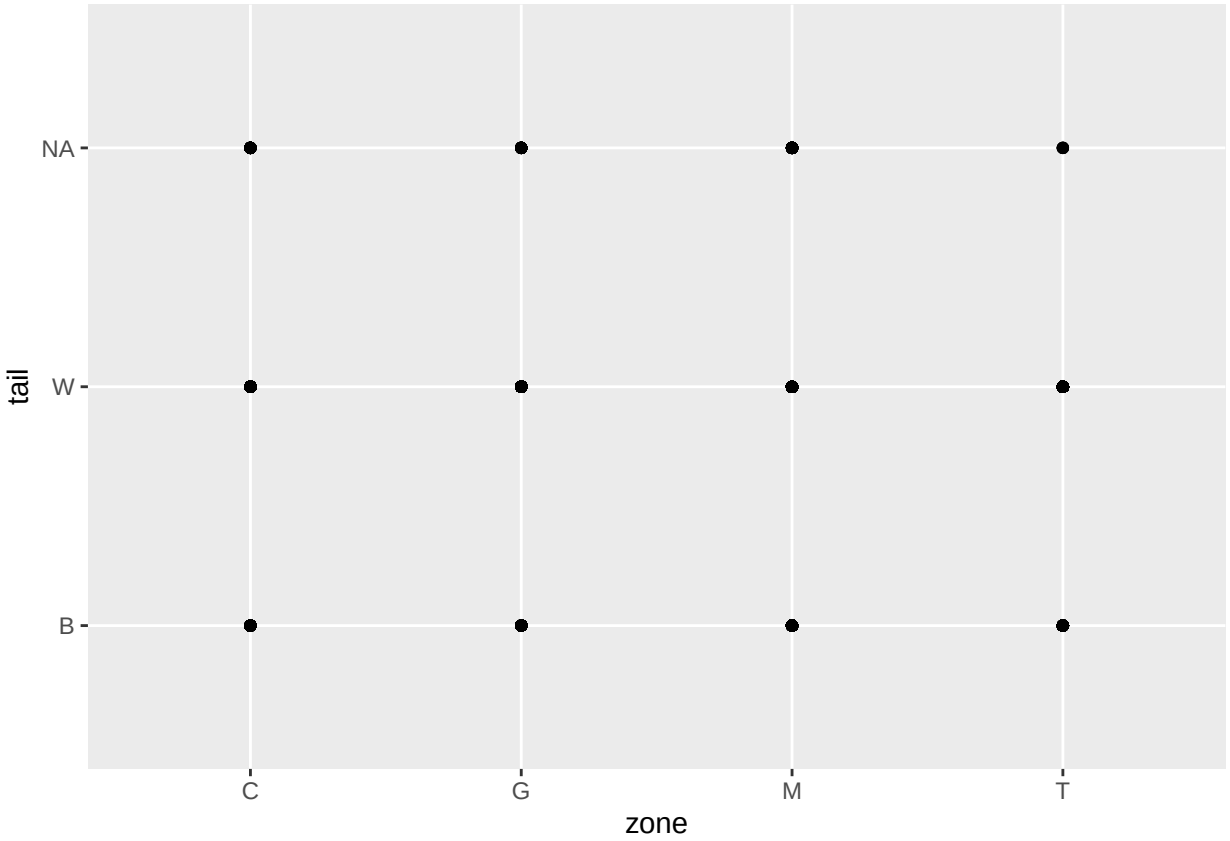

```
ggplot(sb_lizard, aes(x = tail))+
  geom_bar()
```




```
# Relationships with predictor variables
ggplot(sb_lizard, aes(x = sex, y = tail))+
  geom_point()
```

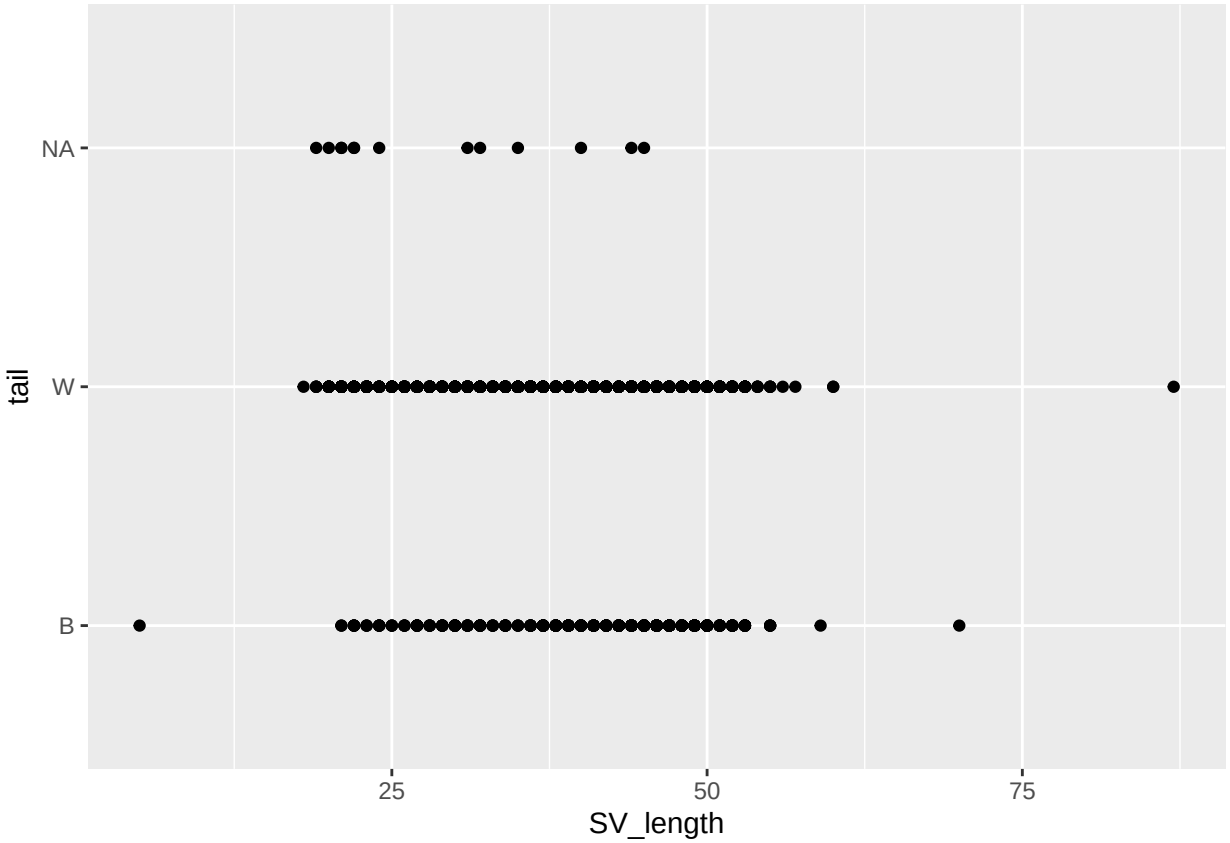



```
ggplot(sb_lizard, aes(x = zone, y = tail))+
  geom_point()
```




```
ggplot(sb_lizard, aes(x = SV_length, y = tail))+
  geom_point()
```

```
## Warning: Removed 114 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
# Set tail to be a factor with a reference level of "W"
sb_lizard$tail <- factor(sb_lizard$tail,
                         levels = c("W", "B"))

# Set sex to be a factor with a reference level of "F"
sb_lizard$sex <- factor(sb_lizard$sex,
                        levels = c("F", "J", "M"))

# Set zone to be a factor with a reference level of "C"
sb_lizard$zone <- factor(sb_lizard$zone,
                         levels = c("C", "G", "M", "T"))

# Logistic regression model
lizard_mod <- glm(tail ~ sex + zone + SV_length,
                  data = sb_lizard,
                  family = "binomial")
```

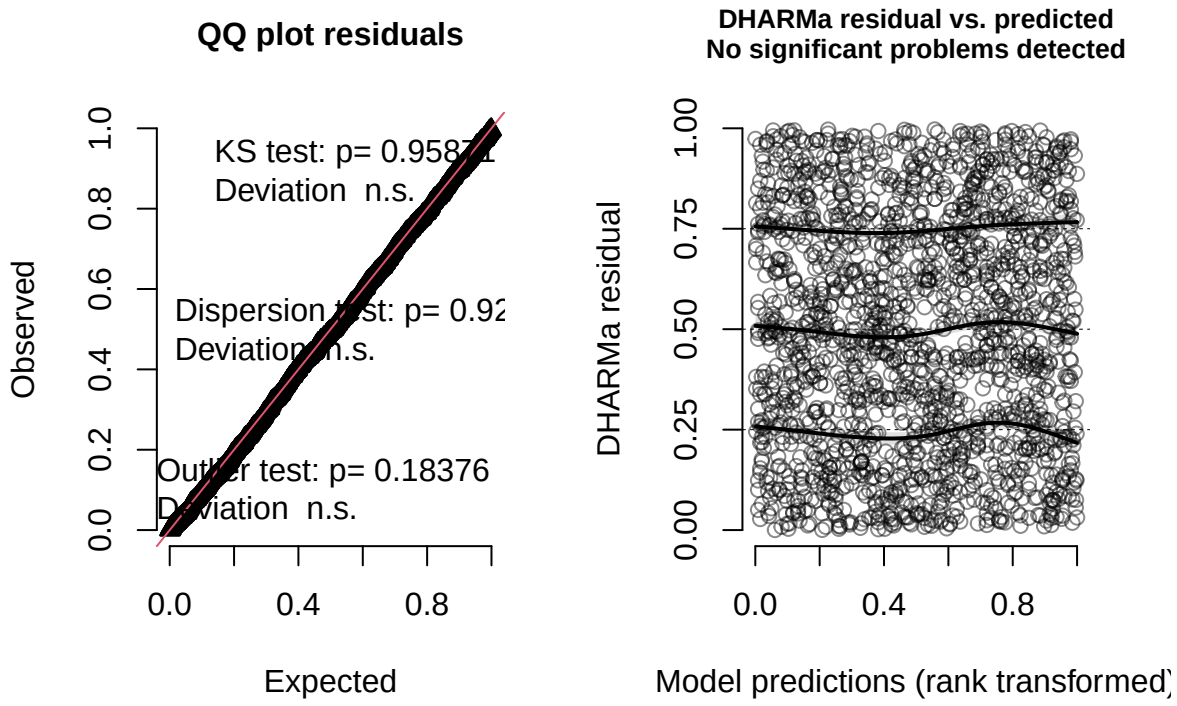
Step 3 - Fit regression model

```
summary(lizard_mod)
```

Step 4 - Evaluate model diagnostics

```
##
## Call:
## glm(formula = tail ~ sex + zone + SV_length, family = "binomial",
##     data = sb_lizard)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.290069   0.407027  -8.083 6.31e-16 ***
## sexJ        -0.203871   0.206033  -0.990  0.32241
## sexM        -0.064915   0.126708  -0.512  0.60842
## zoneG        0.565369   0.173745   3.254  0.00114 **
## zoneM        0.301313   0.162032   1.860  0.06294 .
## zoneT        0.404349   0.179938   2.247  0.02463 *
## SV_length    0.045230   0.008877   5.095 3.49e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1931.7  on 1799  degrees of freedom
## Residual deviance: 1867.3  on 1793  degrees of freedom
##   (138 observations deleted due to missingness)
## AIC: 1881.3
##
## Number of Fisher Scoring iterations: 4
```

```
simulateResiduals(lizard_mod) %>% plot()
```

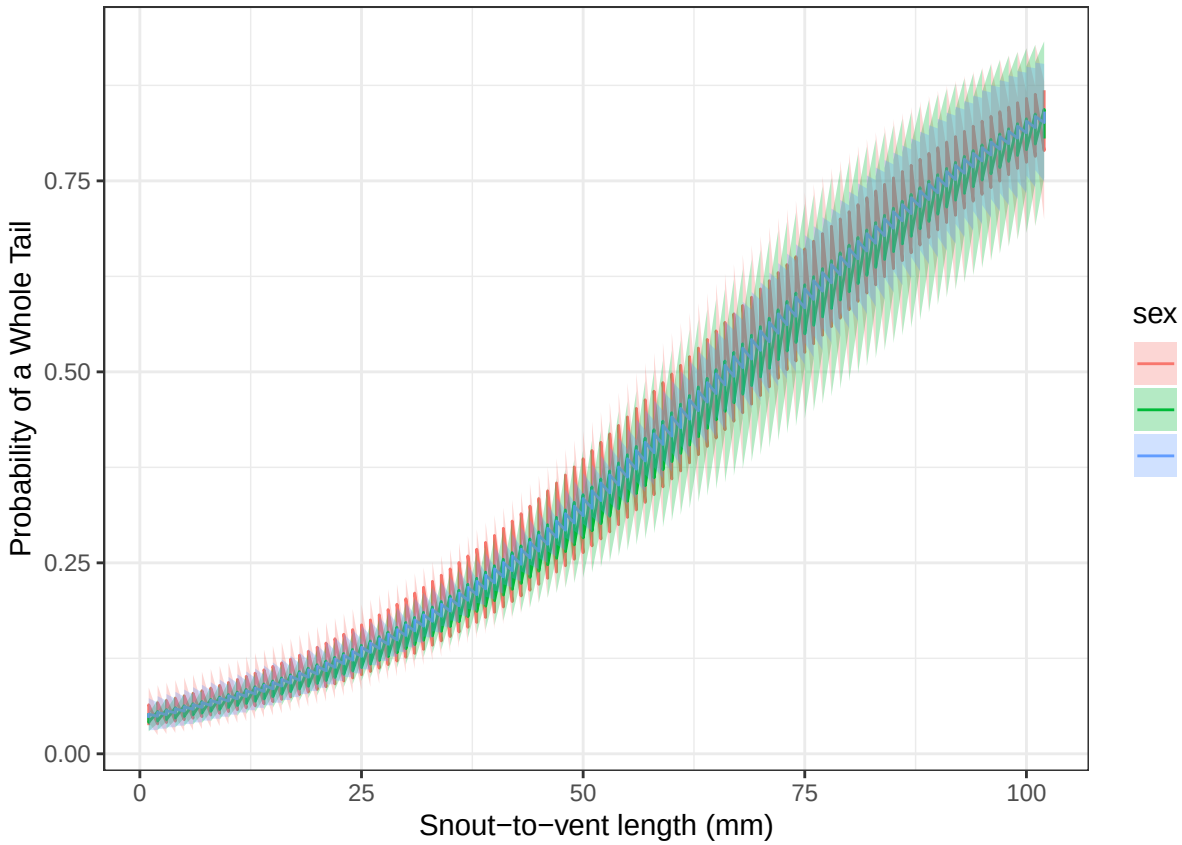


Step 5 - Interpret model results

The results of the logisitic regression indicate that sex did not have significantly different log-odds. The log-odds of tail wholeness increased significantly with grassland (B = 0.565, p = 0.001) and tarbush shrubland (B = 0.404, p = 0.024) in comparison with creosotebush shrubland. Snout-vent-length appeared to increase log-odds of tail wholeness as well (B = 0.045, p < 0.001).

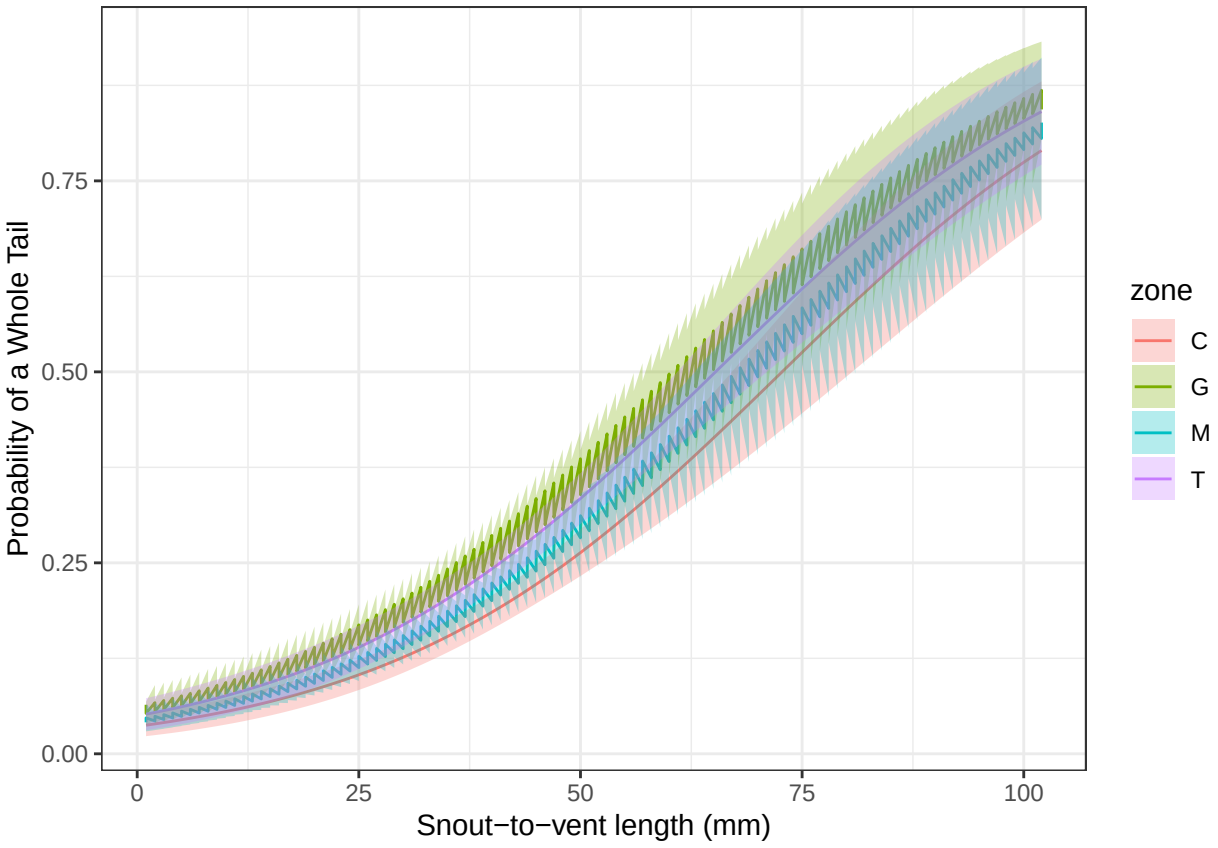
```
# Simulate sex vector
sex_vector <- c(rep("F", 646),
                rep("J", 646),
                rep("M", 646))

# Simulate zone vector
zone_vector <- c(rep("C", 490),
                 rep("G", 490),
                 rep("M", 479),
                 rep("T", 479))

# Simulate SVL vector
SVL_vector <- rep(seq(from = 1, to = 102), 19)
```

```r
# Join data
data_pred <- data.frame(sex_vector, zone_vector, SVL_vector)
colnames(data_pred) <- c("sex", "zone", "SV_length")

# Use original model to predict outcomes
prediction <- predict(lizard_mod,
                      newdata = data_pred,
                      type = "response",
                      se.fit = TRUE)

# Pull out all the predictions
data_fig <- data.frame(data_pred,
                       prediction$fit,
                       prediction$se.fit)

# Rename columns
colnames(data_fig) <- c("sex", "zone", "SV_length", "probability", "se")

# Graph the probabilities of tail wholeness
ggplot(data_fig, aes(x = SV_length,
                     y = probability))+
  geom_line(aes(color = sex))+
  geom_ribbon(aes(ymin = probability - se,
                  ymax = probability + se,
                  fill = sex), alpha = 0.3)+
  labs(x = "Snout-to-vent length (mm)", y = "Probability of a Whole Tail",
       color = "sex", fill = "sex")+
  theme_bw()
```



Predictive plots:

```
ggplot(data_fig, aes(x = SV_length,
                     y = probability))+
  geom_line(aes(color = zone))+
  geom_ribbon(aes(ymin = probability - se,
                  ymax = probability + se,
                  fill = zone), alpha = 0.3)+
  labs(x = "Snout-to-vent length (mm)", y = "Probability of a Whole Tail",
       color = "zone", fill = "zone")+
  theme_bw()
```



(2) Poisson Regression

```
# Load in data
npp_lizard <- read_csv("jrn_lizard_npp.csv")
```

```
## Rows: 52 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): season
## dbl (6): sample_year, lizard_count, BOER, LATR, PRGL, SCBR
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

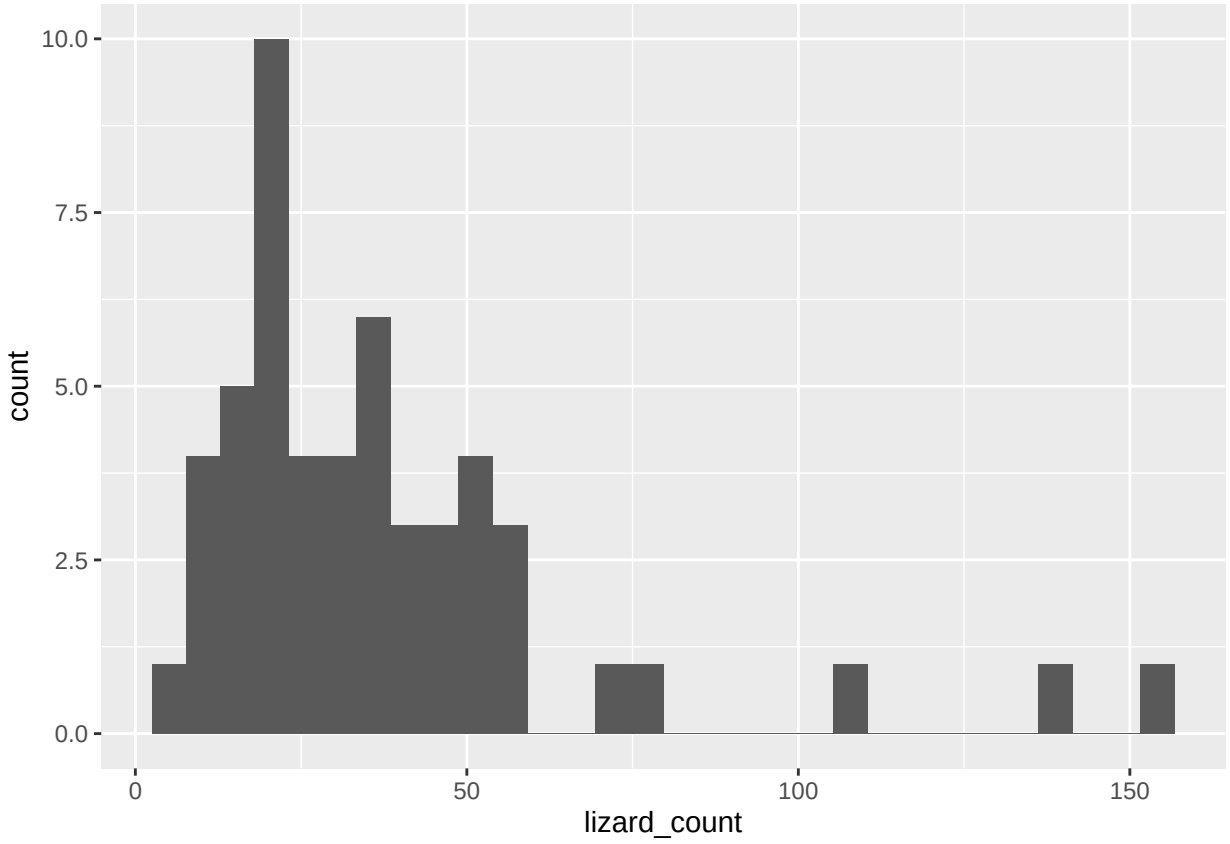
Step 1: Define the research question

"Do season and plant species percent cover significantly predict lizard counts?

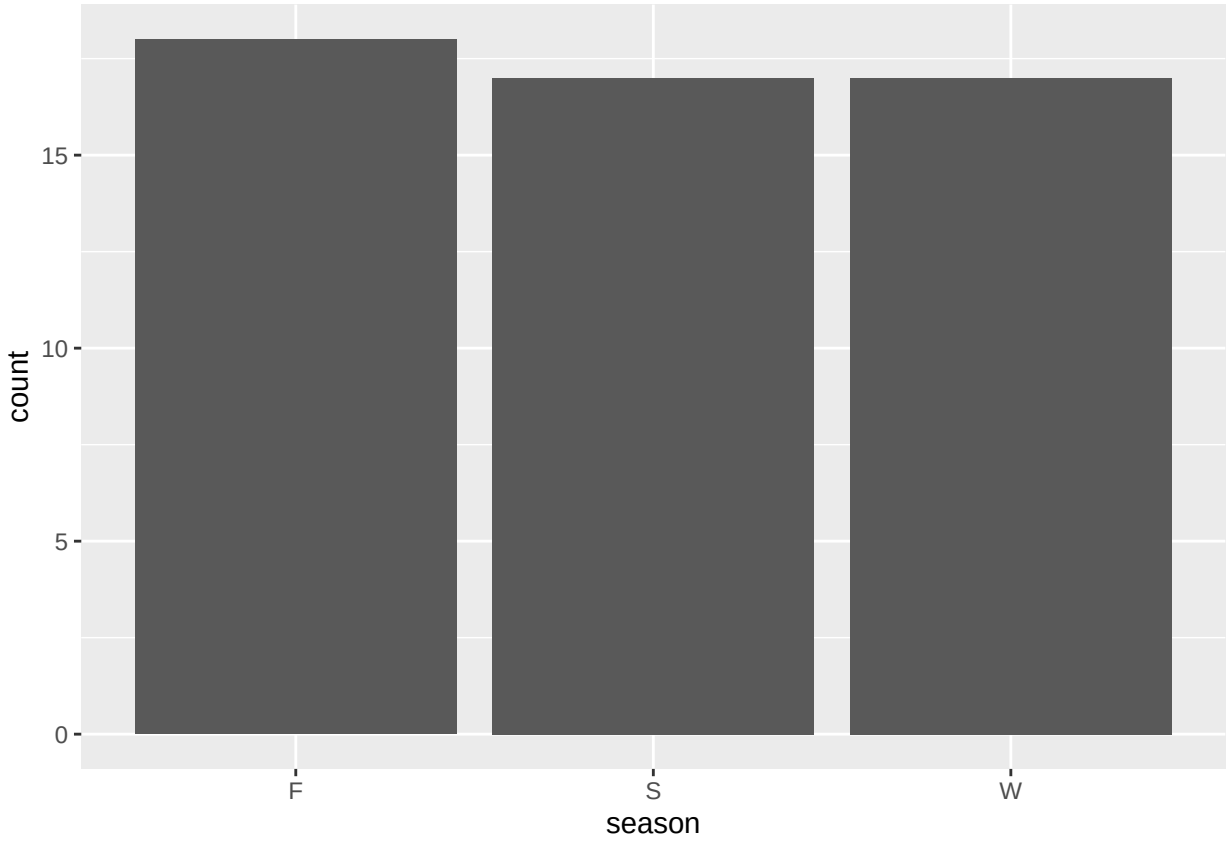
```
# Raw counts
ggplot(npp_lizard, aes(x = lizard_count))+
  geom_histogram() # right skewed
```

Step 2: Examine data and possible correlations

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

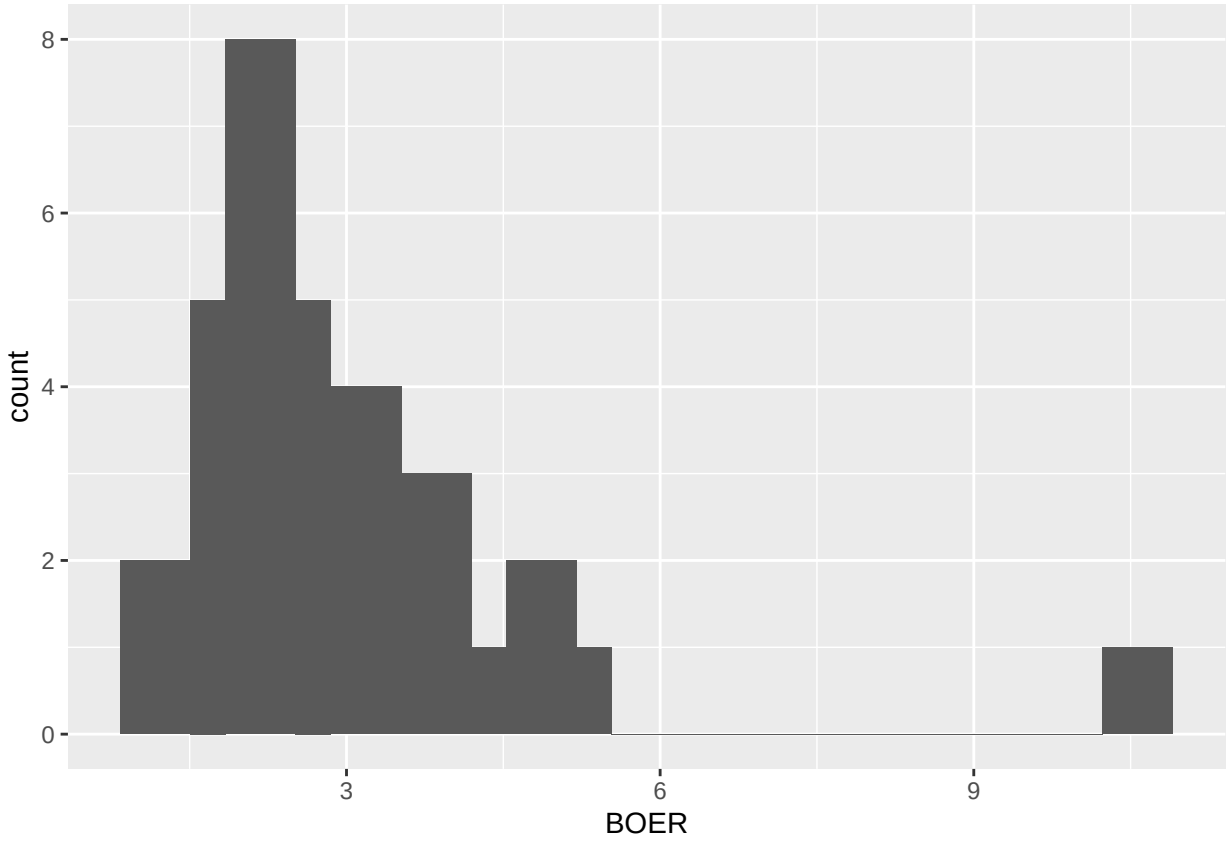



```
ggplot(npp_lizard, aes(x = season))+
  geom_bar()
```



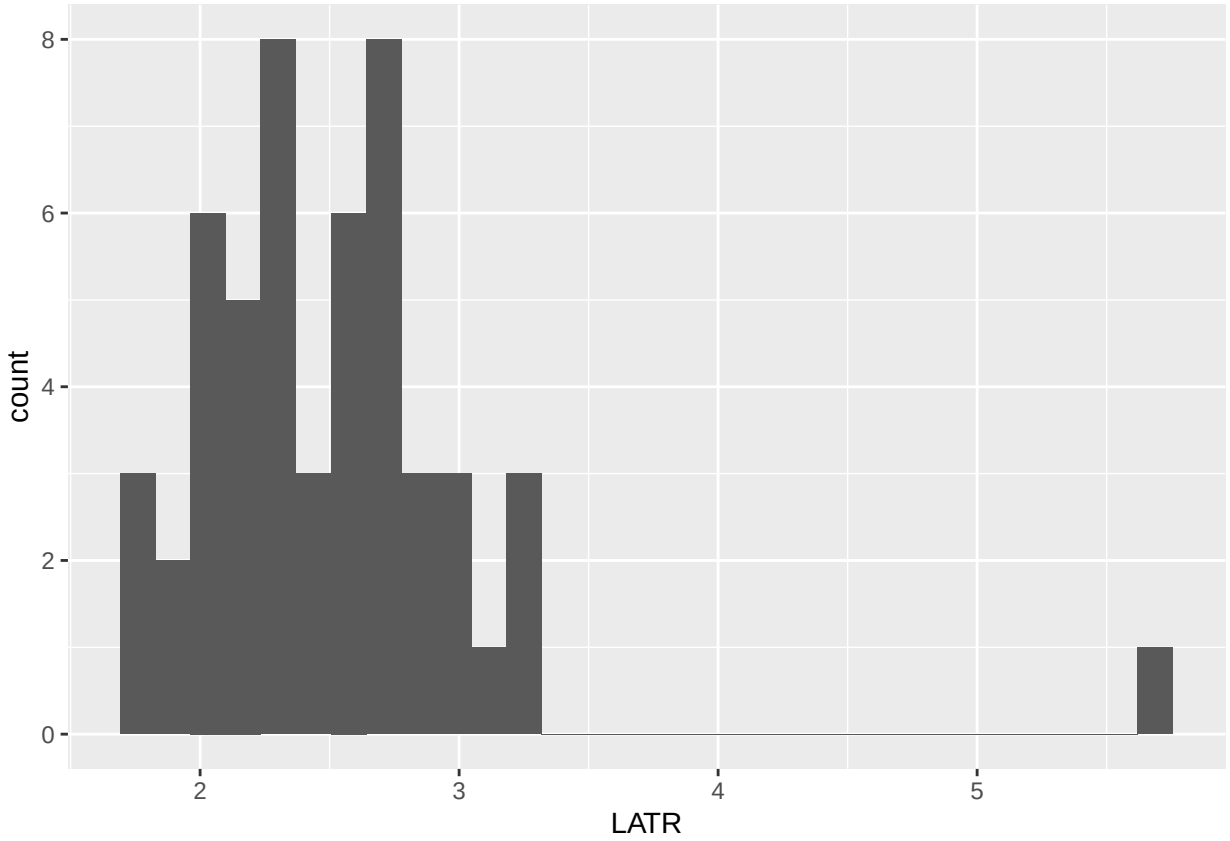

```
ggplot(npp_lizard, aes(x = BOER))+
  geom_histogram() # possible outlier
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



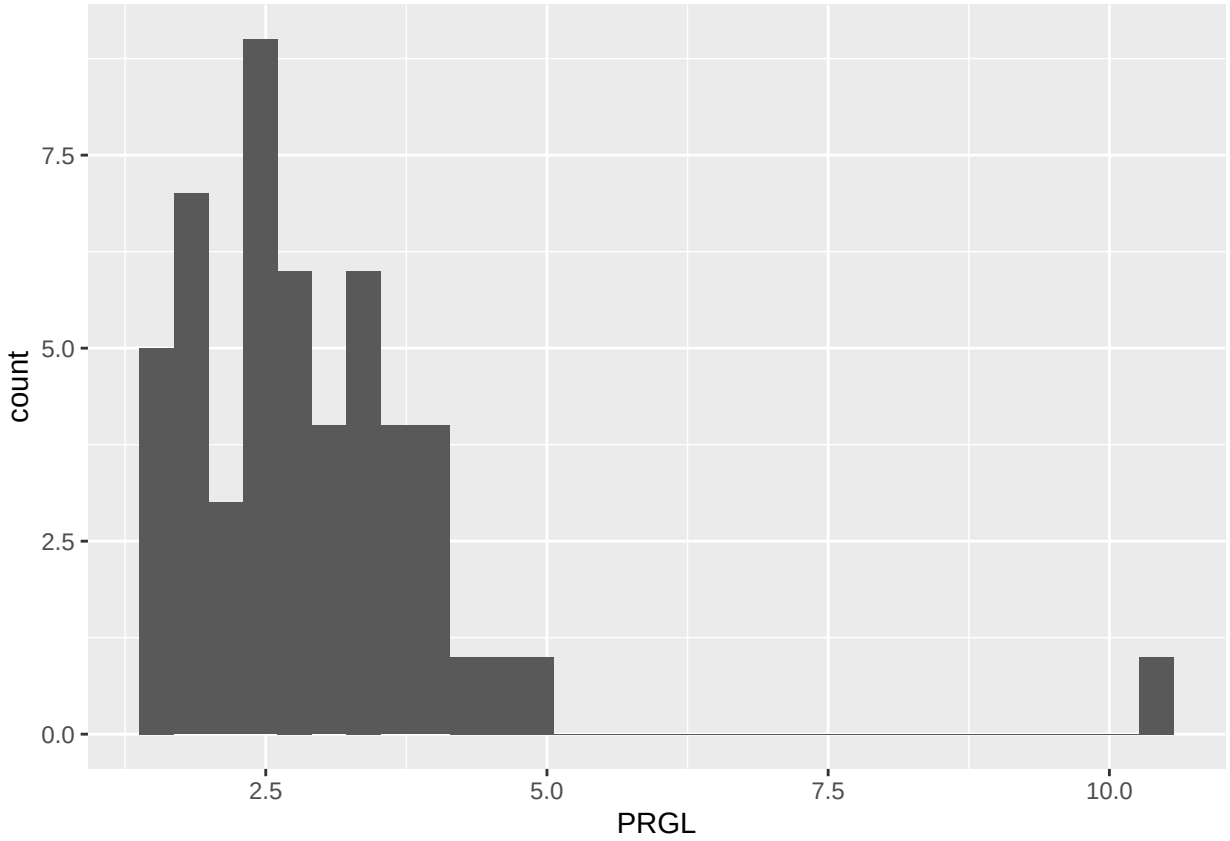

```
ggplot(npp_lizard, aes(x = LATR))+
  geom_histogram() # possible outlier
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(npp_lizard, aes(x = PRGL))+
  geom_histogram() # possible outlier
```

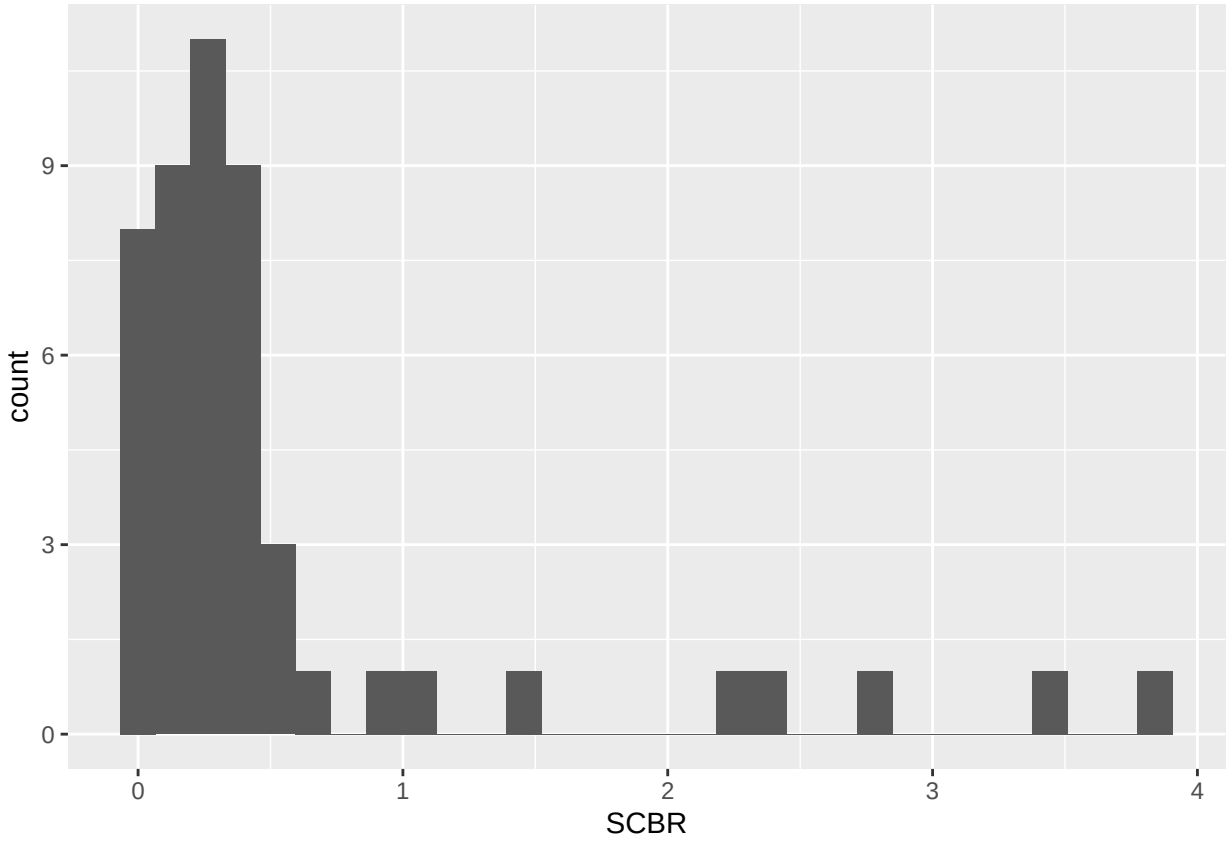
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



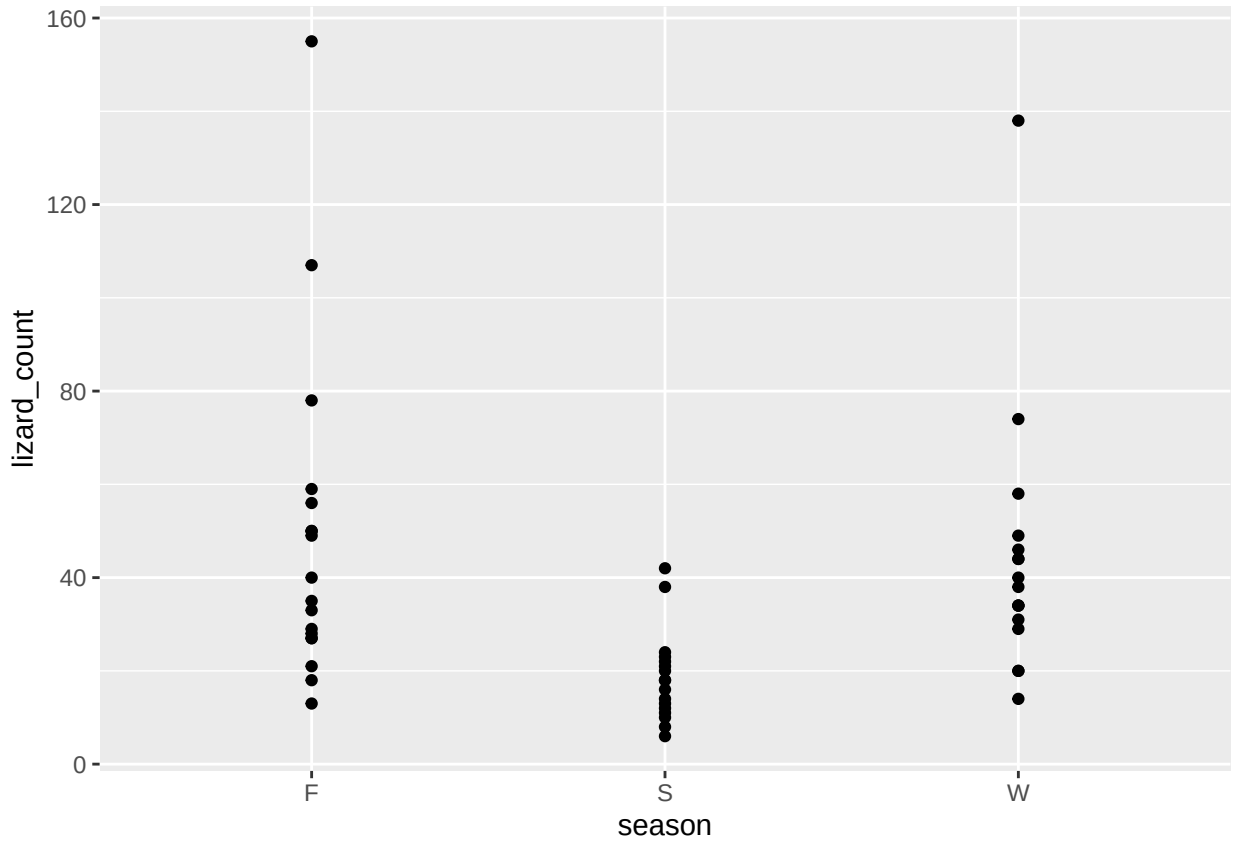

```
ggplot(npp_lizard, aes(x = SCBR))+
  geom_histogram() # right skewed
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

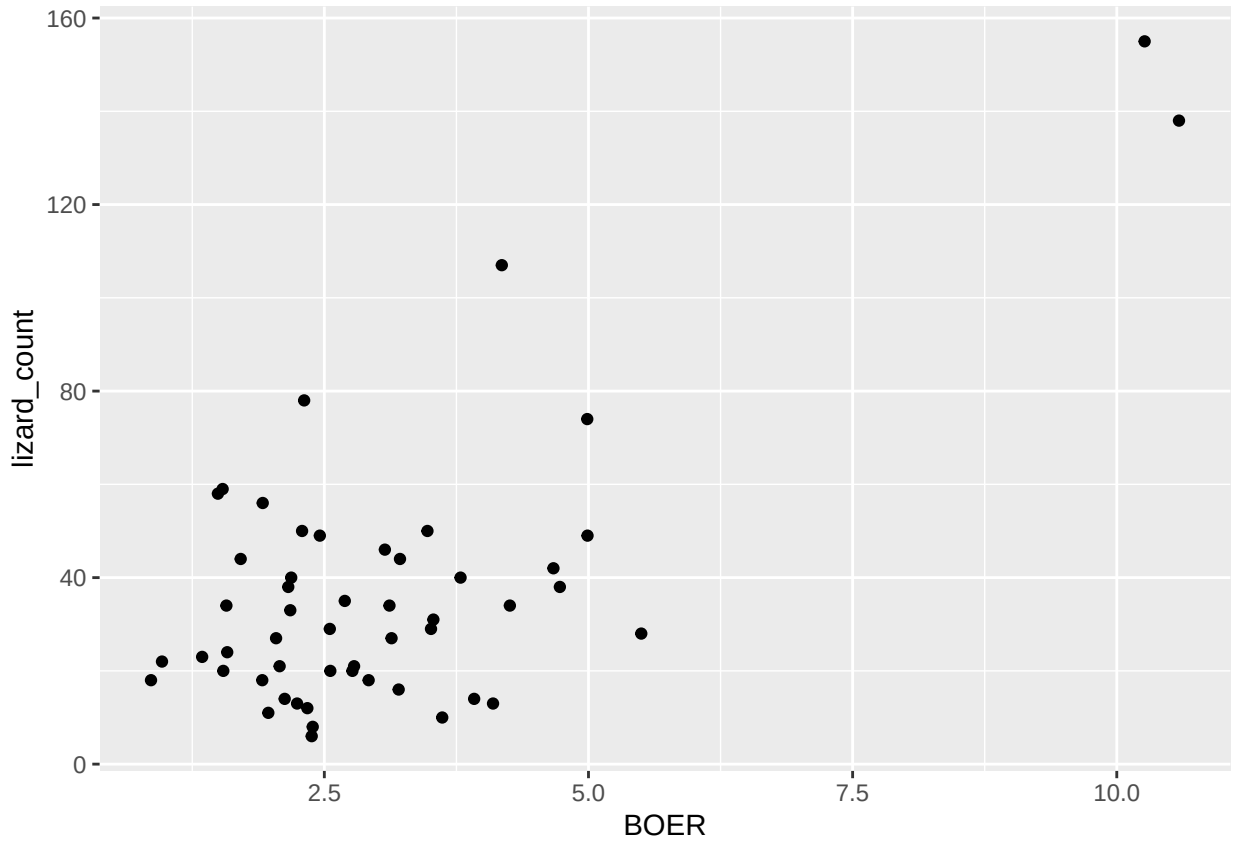
```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



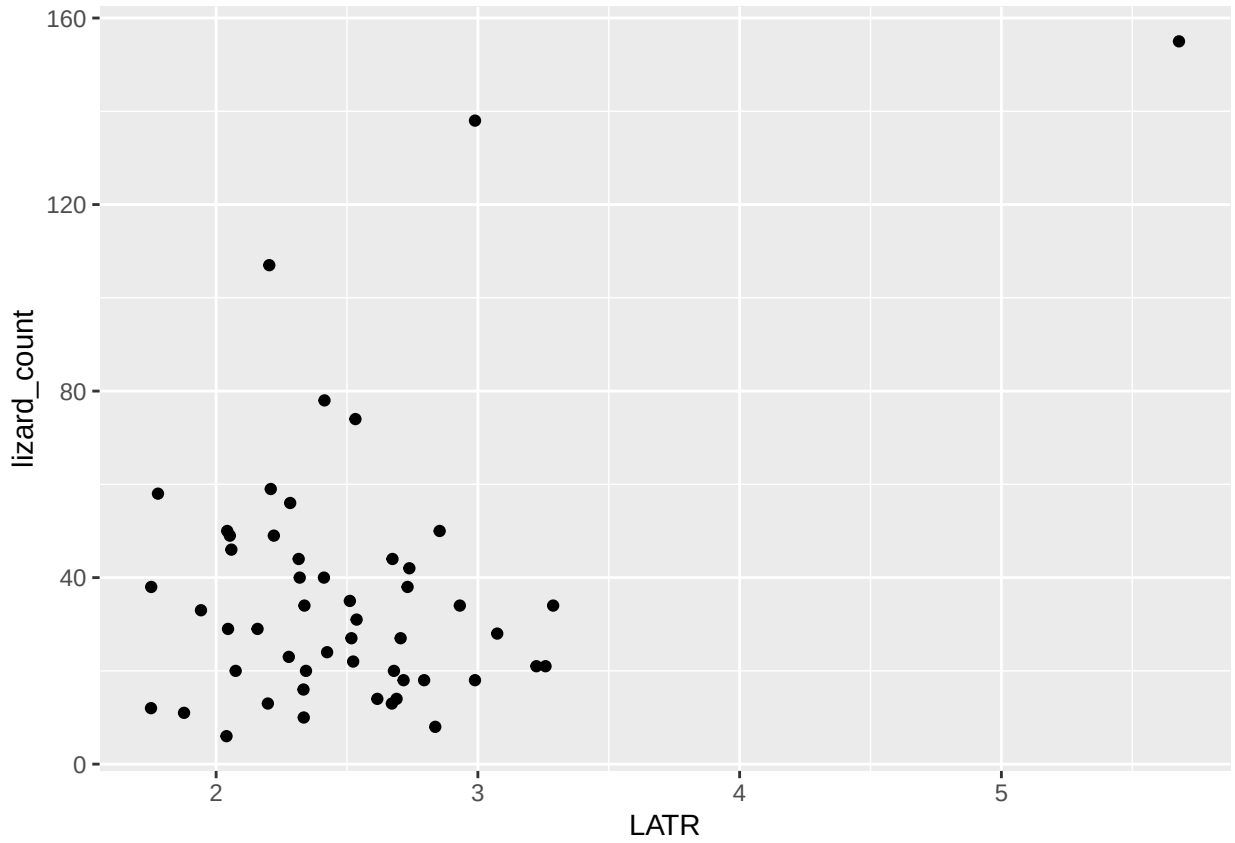

```
# Relationships with predictor variables
ggplot(npp_lizard, aes(x = season, y = lizard_count))+
  geom_point() # less in summer
```



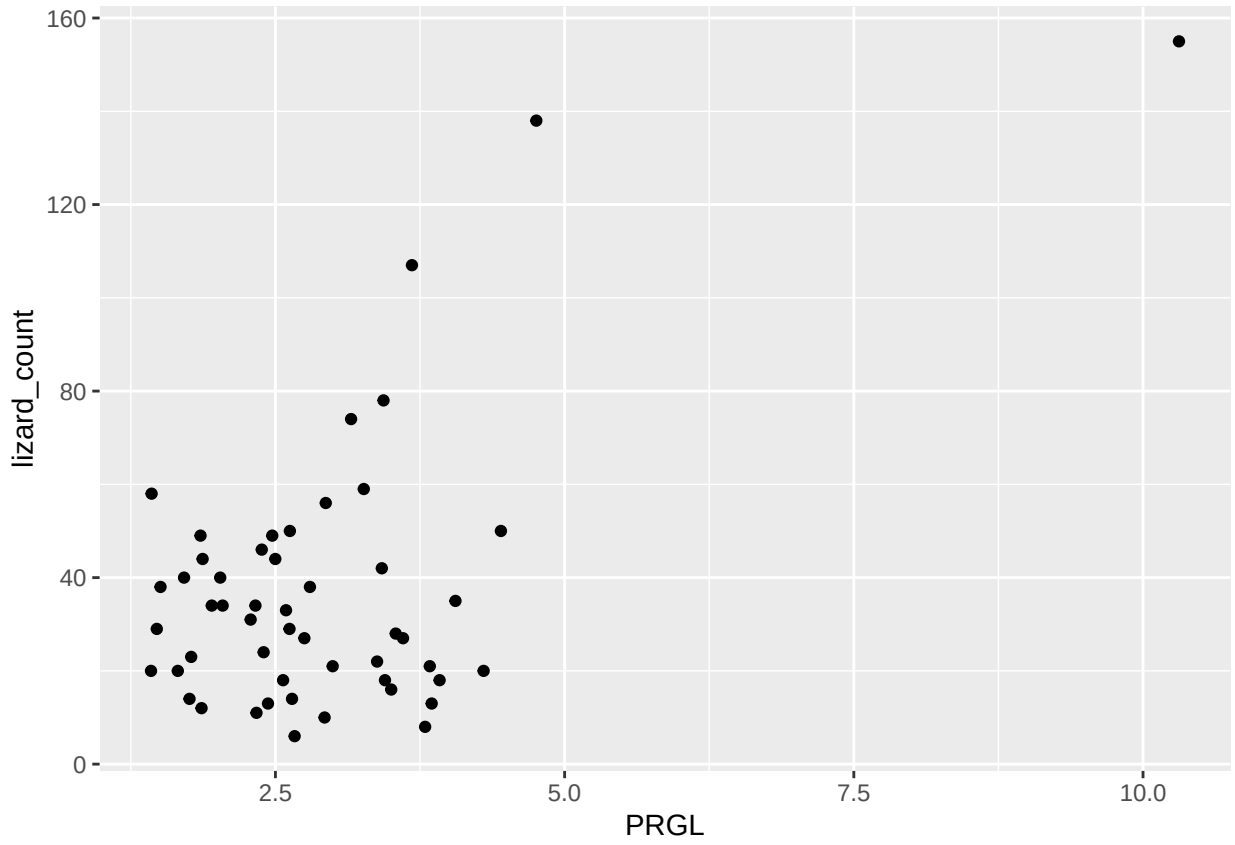

```
ggplot(npp_lizard, aes(x = BOER, y = lizard_count))+
  geom_point() # positive linear
```




```
ggplot(npp_lizard, aes(x = LATR, y = lizard_count))+
  geom_point() # positive linear
```

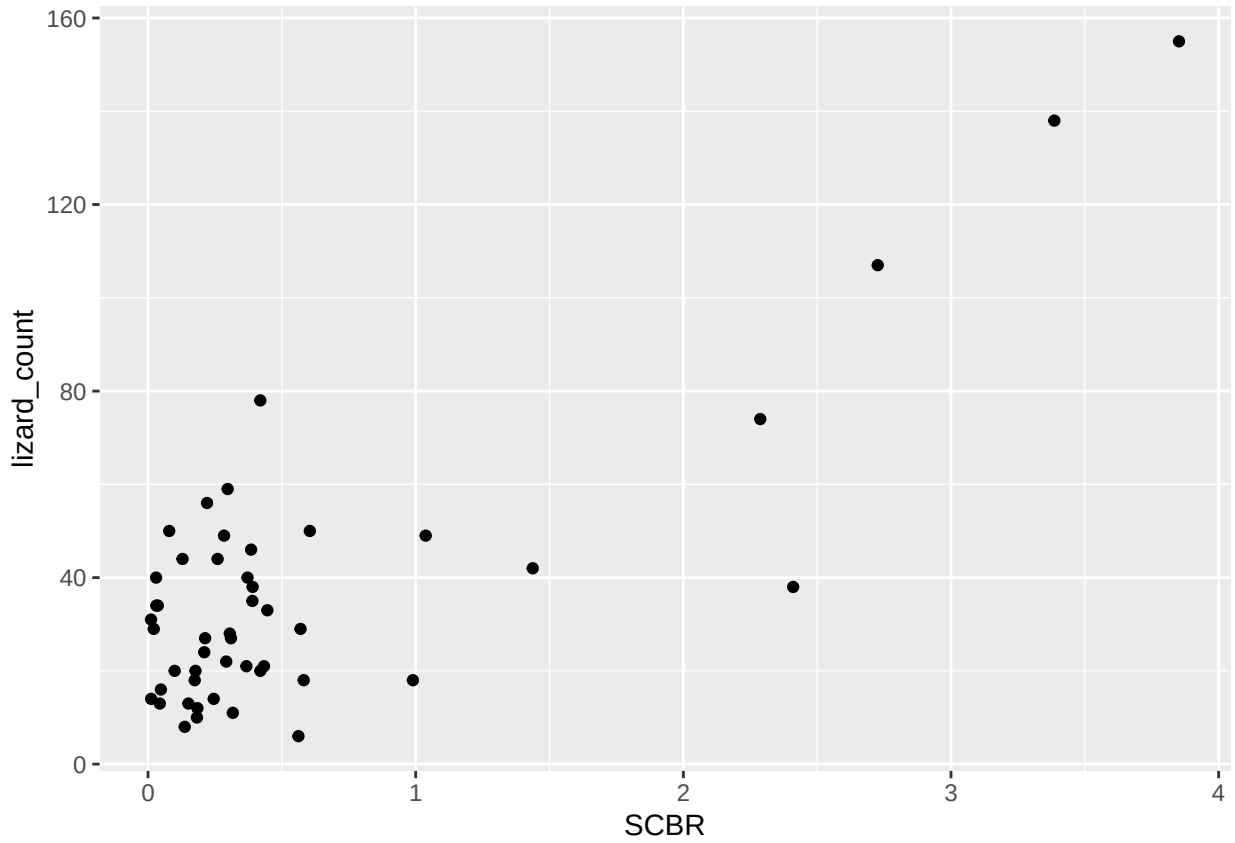



```
ggplot(npp_lizard, aes(x = PRGL, y = lizard_count))+
  geom_point() # positive linear
```




```
ggplot(npp_lizard, aes(x = SCBR, y = lizard_count))+
  geom_point() # positive linear
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
# It looks like lizards are less common in the summer; include BOER and SCBR,
# may have stronger influence on counts
```

```
# Set season as a factor with a reference level of "F"
npp_lizard$season <- factor(npp_lizard$season,
                            levels = c("F", "S", "W"))

# Poisson regression model
npp_lizard_mod <- glm(lizard_count ~ season + BOER + SCBR,
                      data = npp_lizard,
                      family = "poisson")
```

Step 3: Fit regression model

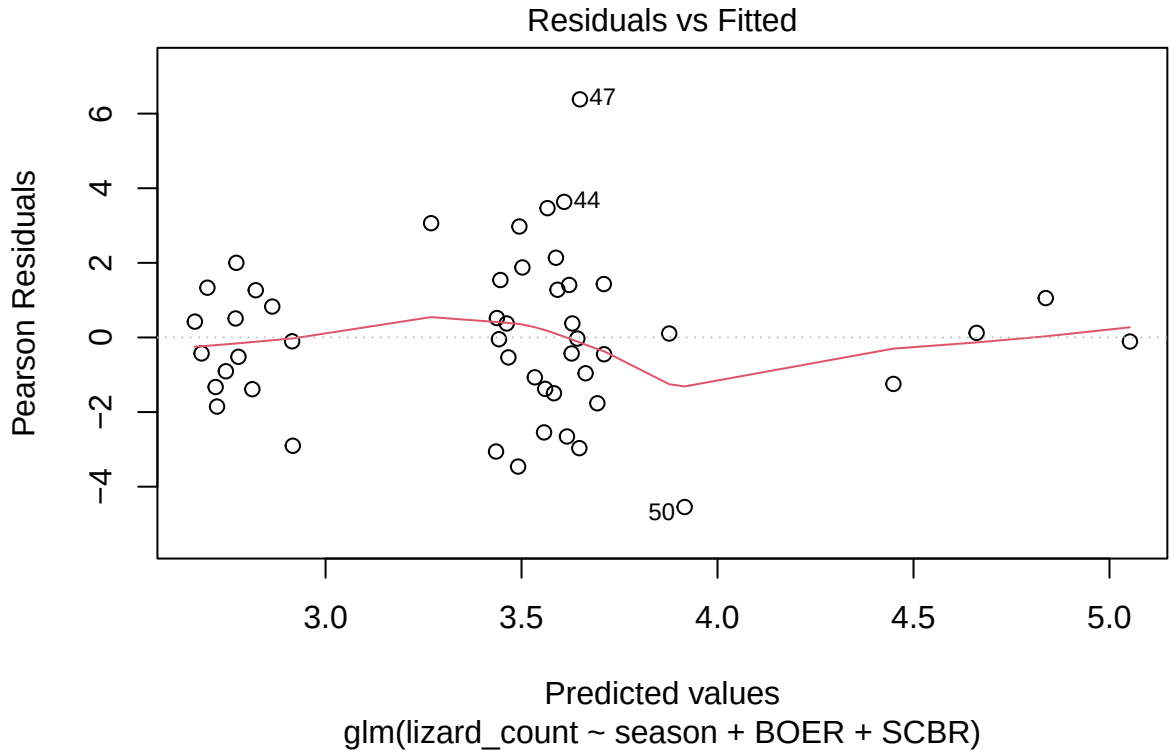
```
summary(npp_lizard_mod)
```

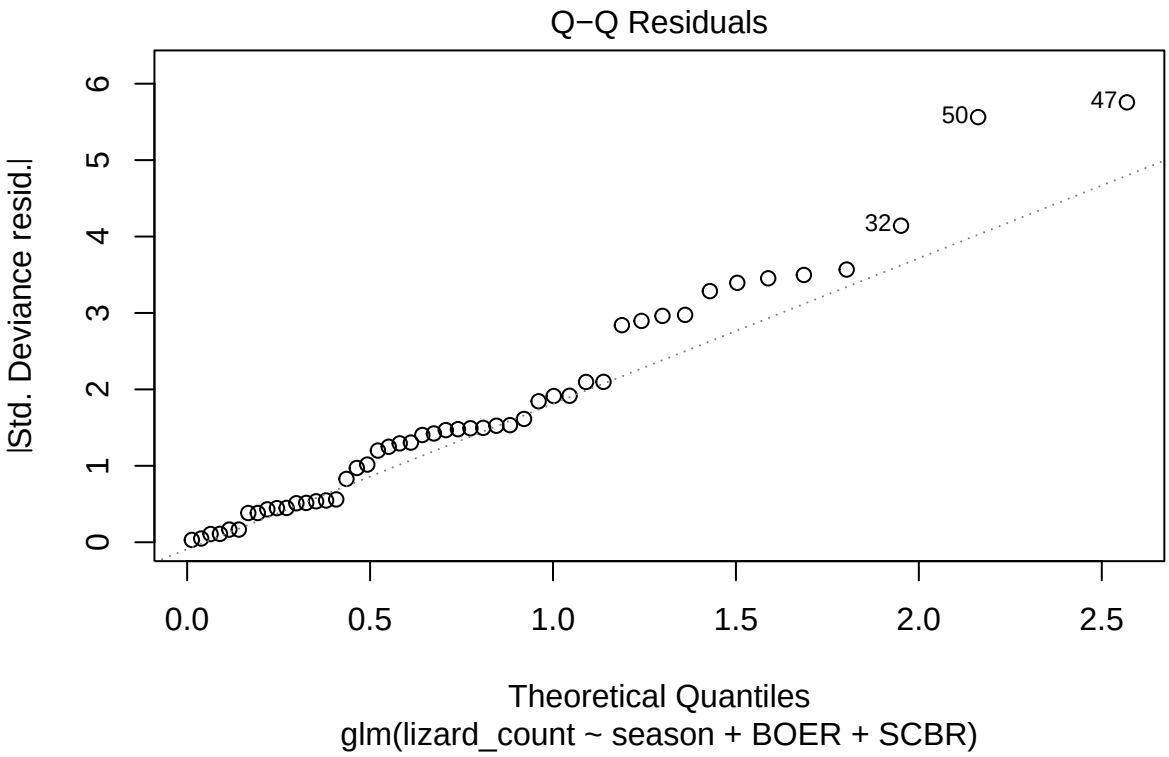
Step 4: Evaluate model diagnostics

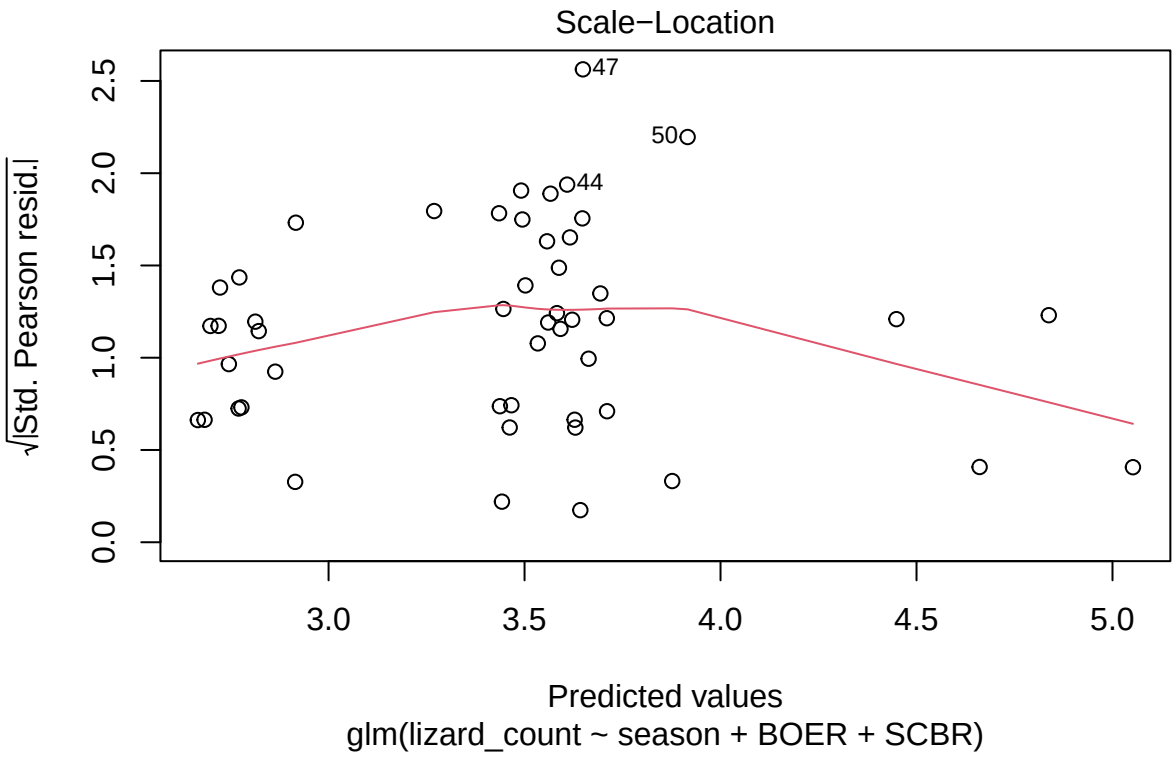
##

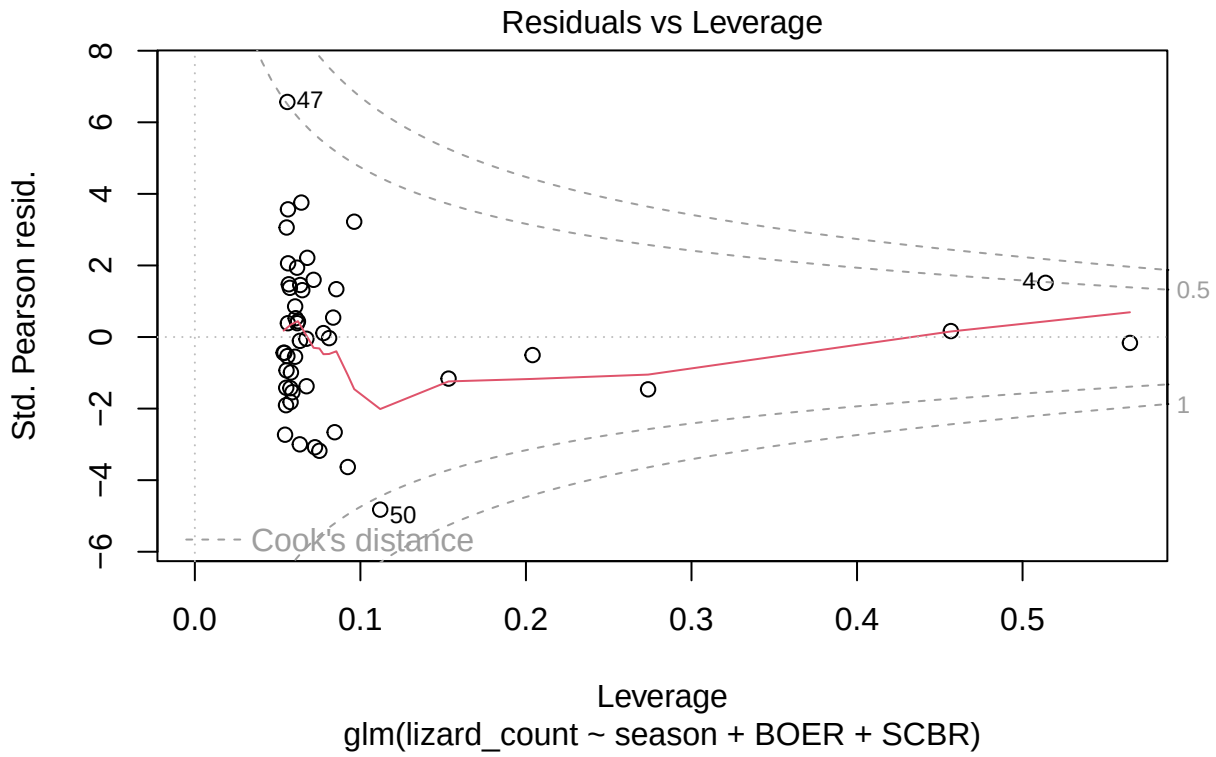
```
## Call:
## glm(formula = lizard_count ~ season + BOER + SCBR, family = "poisson",
##     data = npp_lizard)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.50388    0.05314  65.933   <2e-16 ***
## seasonS     -0.79572    0.06881 -11.565   <2e-16 ***
## seasonW      0.00349    0.05357   0.065    0.948
## BOER        -0.01983    0.01783  -1.112    0.266
## SCBR         0.45480    0.03822  11.898   <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: 911.44  on 48  degrees of freedom
## Residual deviance: 207.41  on 44  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 473.07
##
## Number of Fisher Scoring iterations: 4
```

```
plot(npp_lizard_mod)
```




Q−Q Residuals
|Std. Deviance resid.|
0
1
2
3
4
5
6
0.0
0.5
1.0
1.5
2.0
2.5
47
50
32
Theoretical Quantiles
glm(lizard_count ~ season + BOER + SCBR)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
2.5
47
50
44
3.0
3.5
4.0
4.5
5.0
Predicted values
glm(lizard_count ~ season + BOER + SCBR)

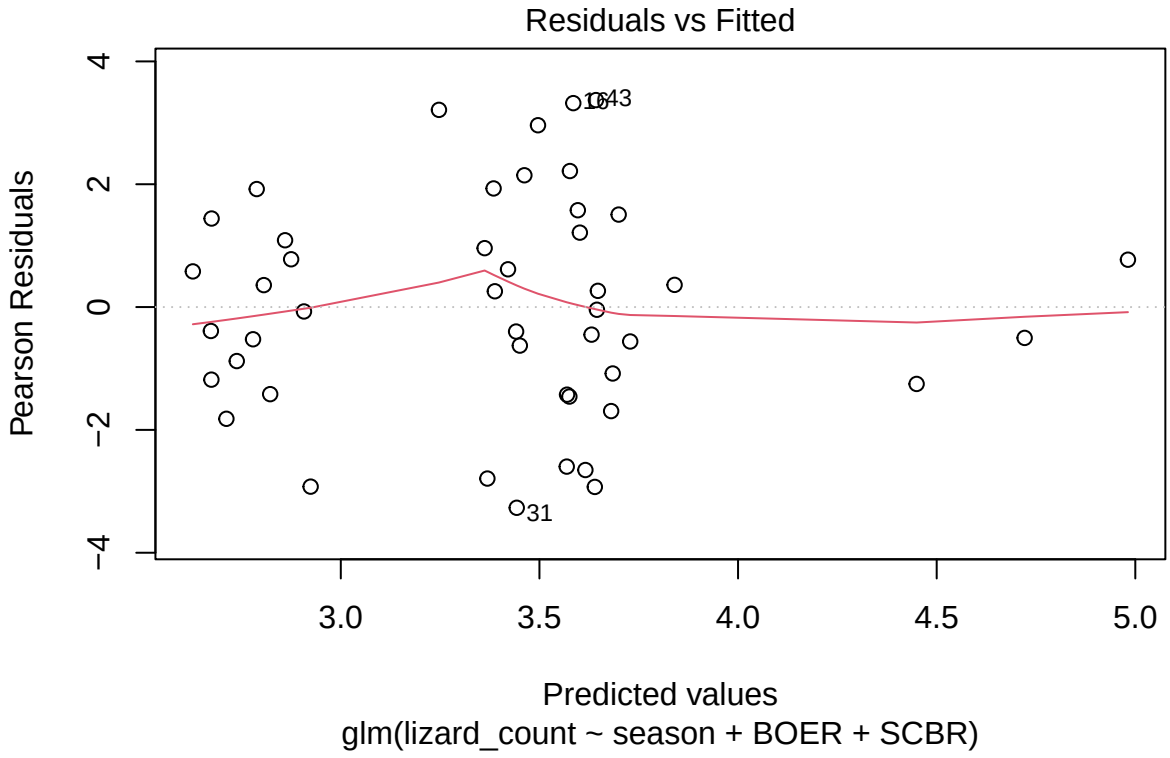


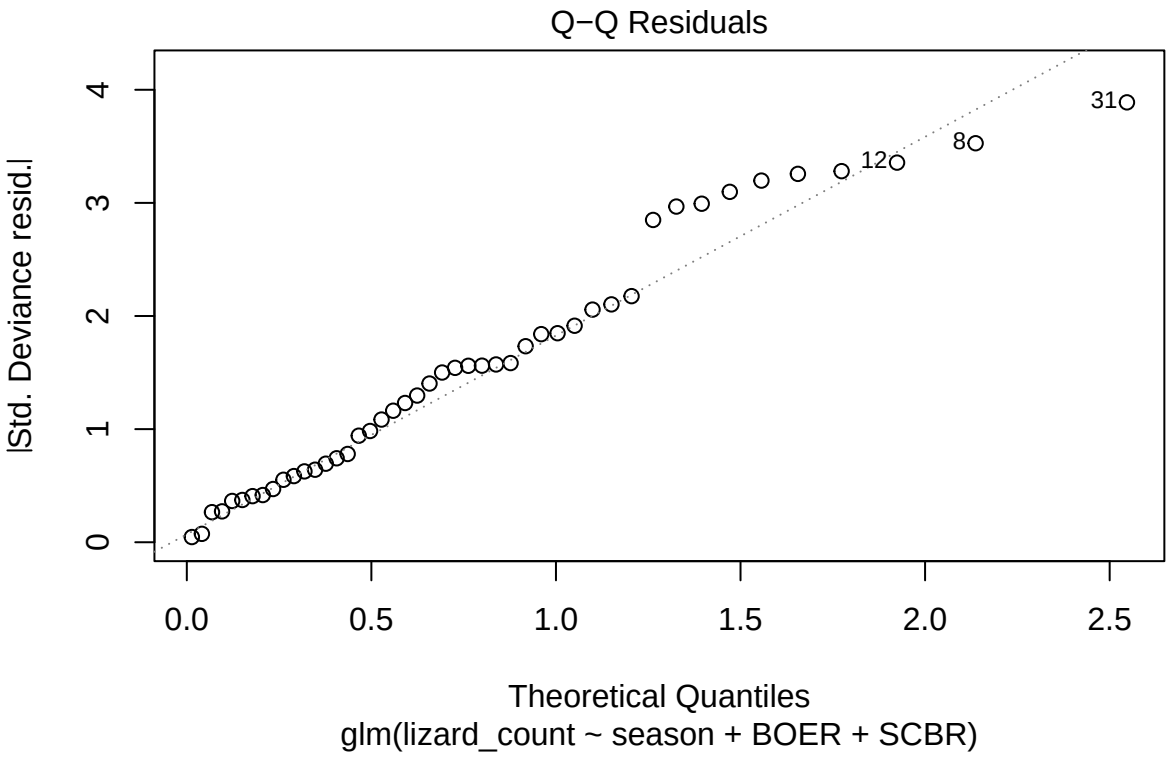

```
# Outliers include 50, 47, 4

# Remove outliers
npp_lizard2 <- npp_lizard[-c(4, 47, 50),]

# Refit the model
npp_lizard_mod2 <- glm(lizard_count ~ season + BOER + SCBR,
                       data = npp_lizard2,
                       family = "poisson")

# Re-examine model results
summary(npp_lizard_mod2)
```

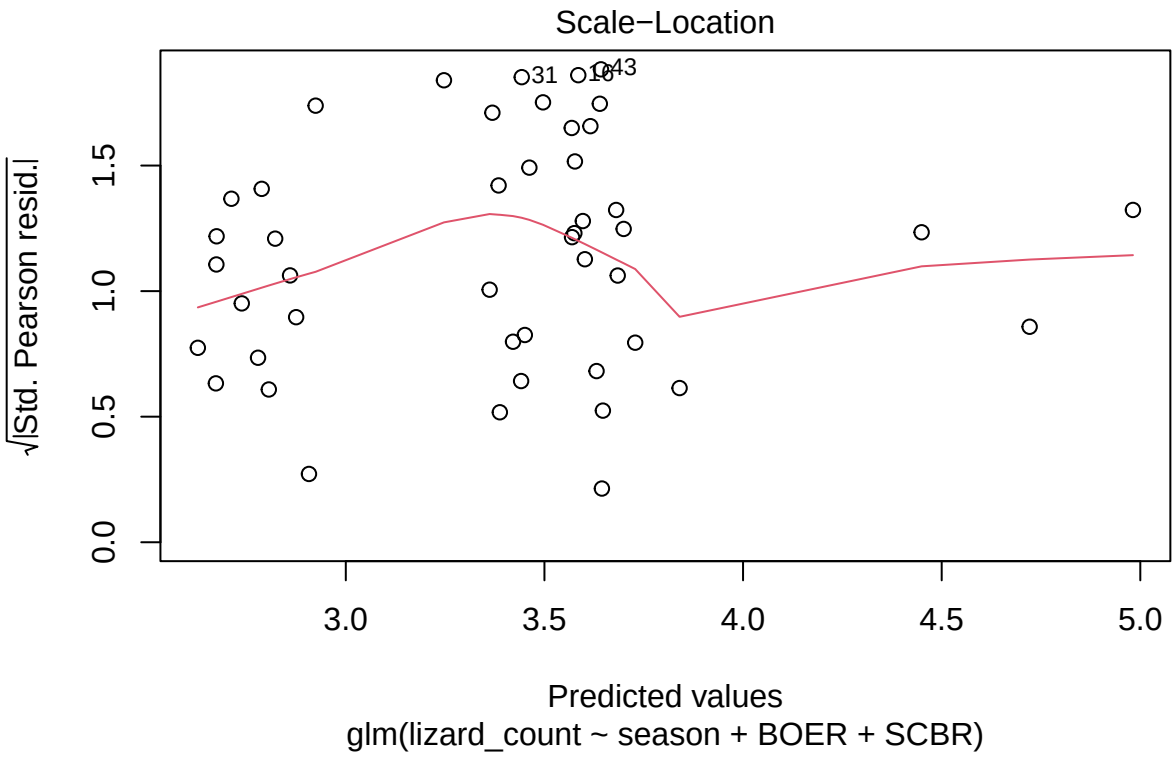
```
##
## Call:
## glm(formula = lizard_count ~ season + BOER + SCBR, family = "poisson",
##     data = npp_lizard2)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.57001    0.06735  53.009   <2e-16 ***
## seasonS     -0.80827    0.07144 -11.314   <2e-16 ***
## seasonW     -0.01335    0.05965  -0.224   0.8229
## BOER        -0.04937    0.02059  -2.398   0.0165 *
## SCBR         0.49799    0.03913  12.727   <2e-16 ***
## ---
```

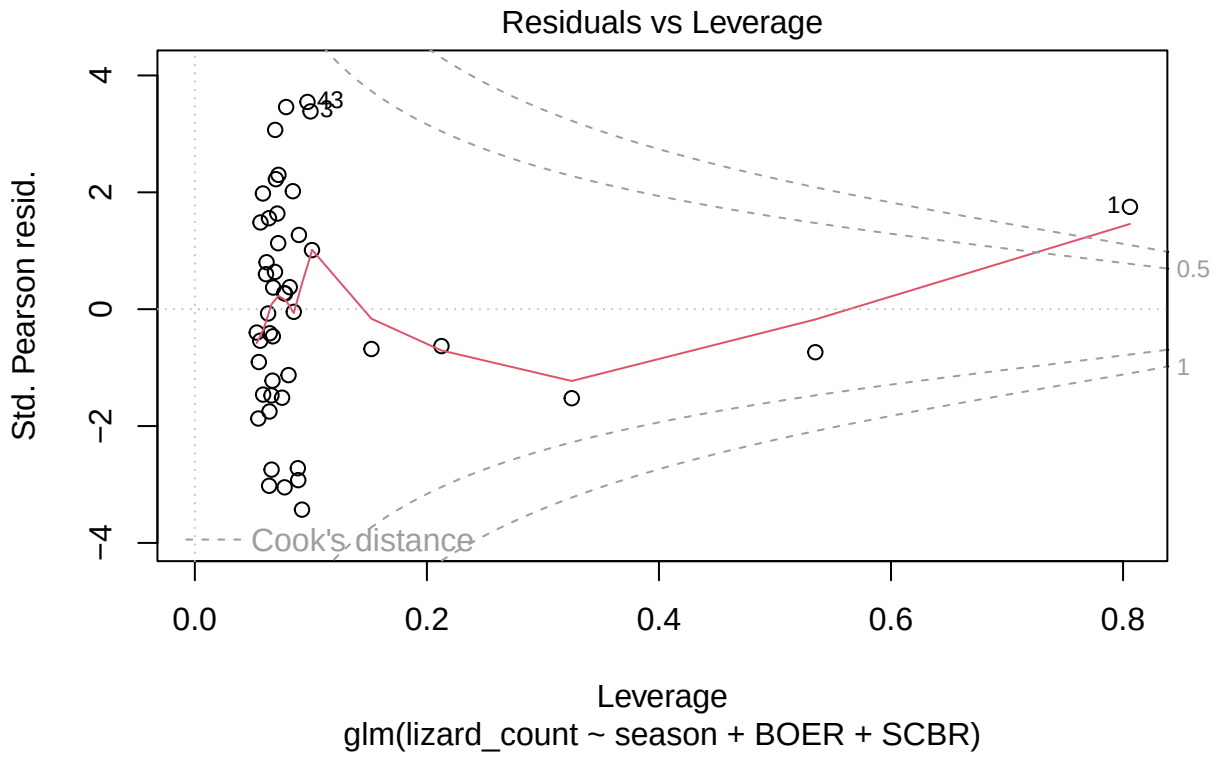
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 696.10  on 45  degrees of freedom
## Residual deviance: 145.14  on 41  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 393.1
##
## Number of Fisher Scoring iterations: 4
```

```
plot(npp_lizard_mod2)
```




Q−Q Residuals
|Std. Deviance resid.|
0
1
2
3
4
0.0
0.5
1.0
1.5
2.0
2.5
31
8
12
Theoretical Quantiles
glm(lizard_count ~ season + BOER + SCBR)


Scale−Location
31
43
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
3.0
3.5
4.0
4.5
5.0
Predicted values
glm(lizard_count ~ season + BOER + SCBR)

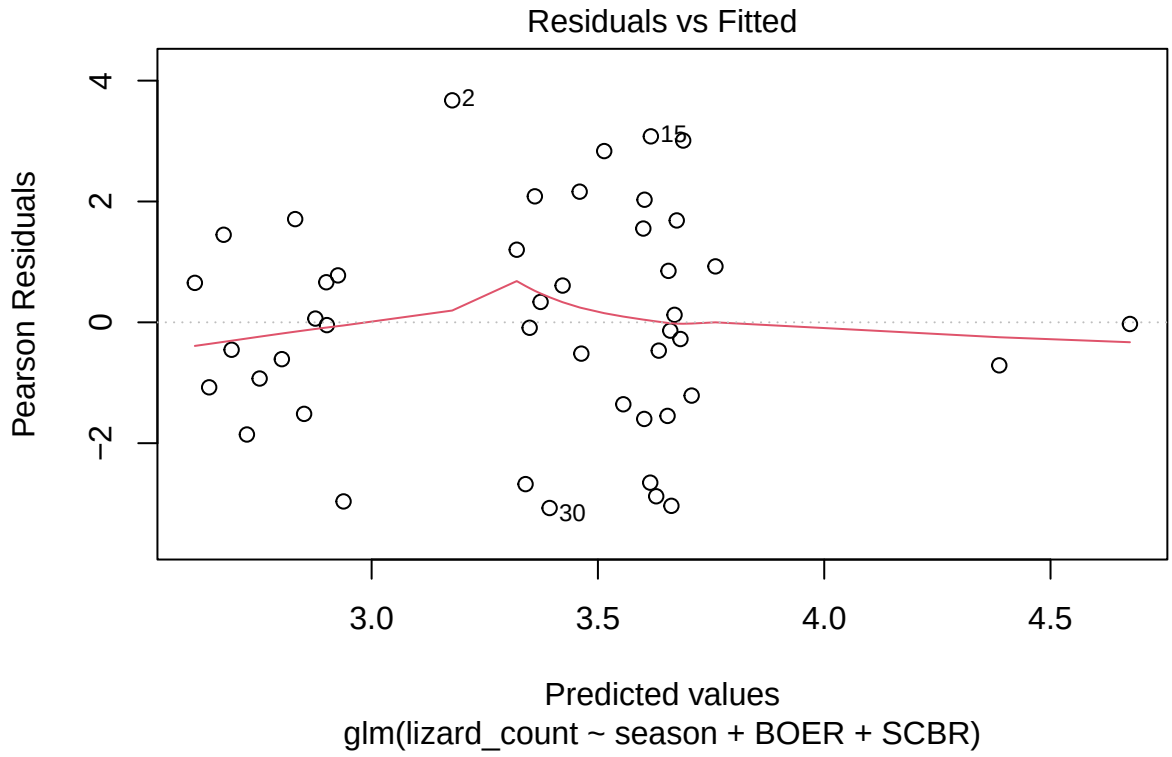


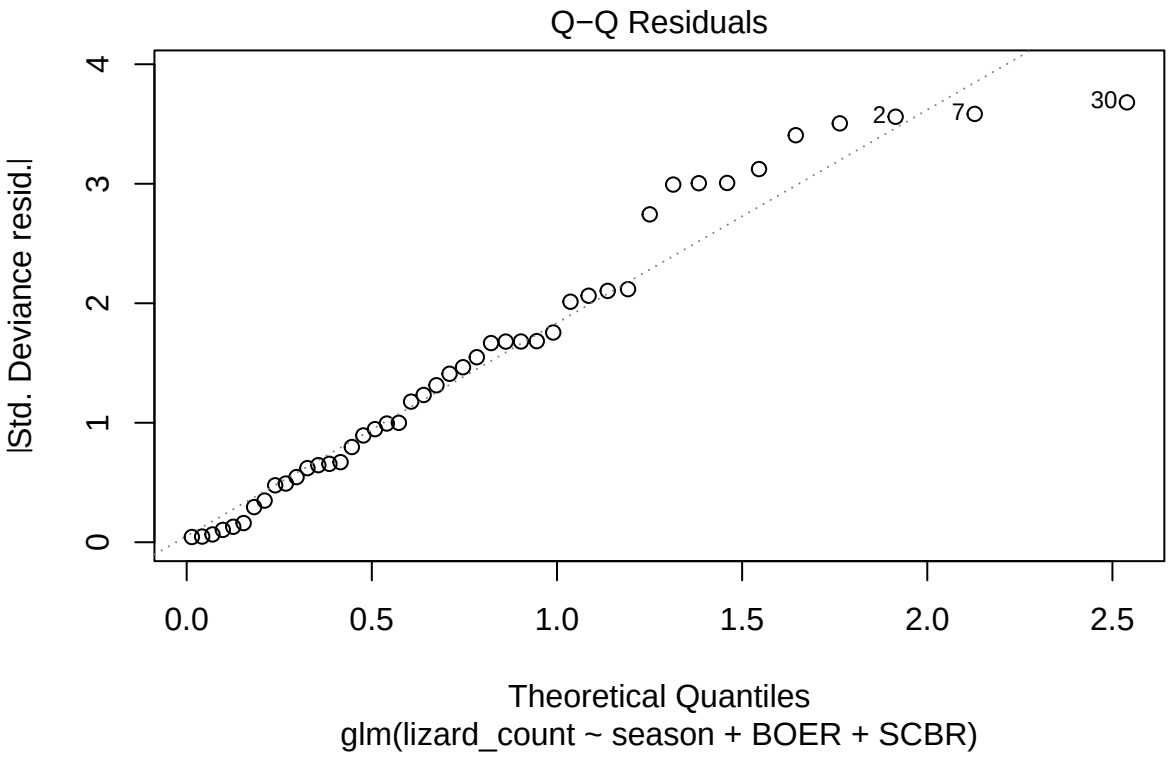

```
# Remove 1 more outlier
npp_lizard3 <- npp_lizard2[-1,]

# Refit the model
npp_lizard_mod3 <- glm(lizard_count ~ season + BOER + SCBR,
                       data = npp_lizard3,
                       family = "poisson")

# Re-examine model results
summary(npp_lizard_mod3)
```

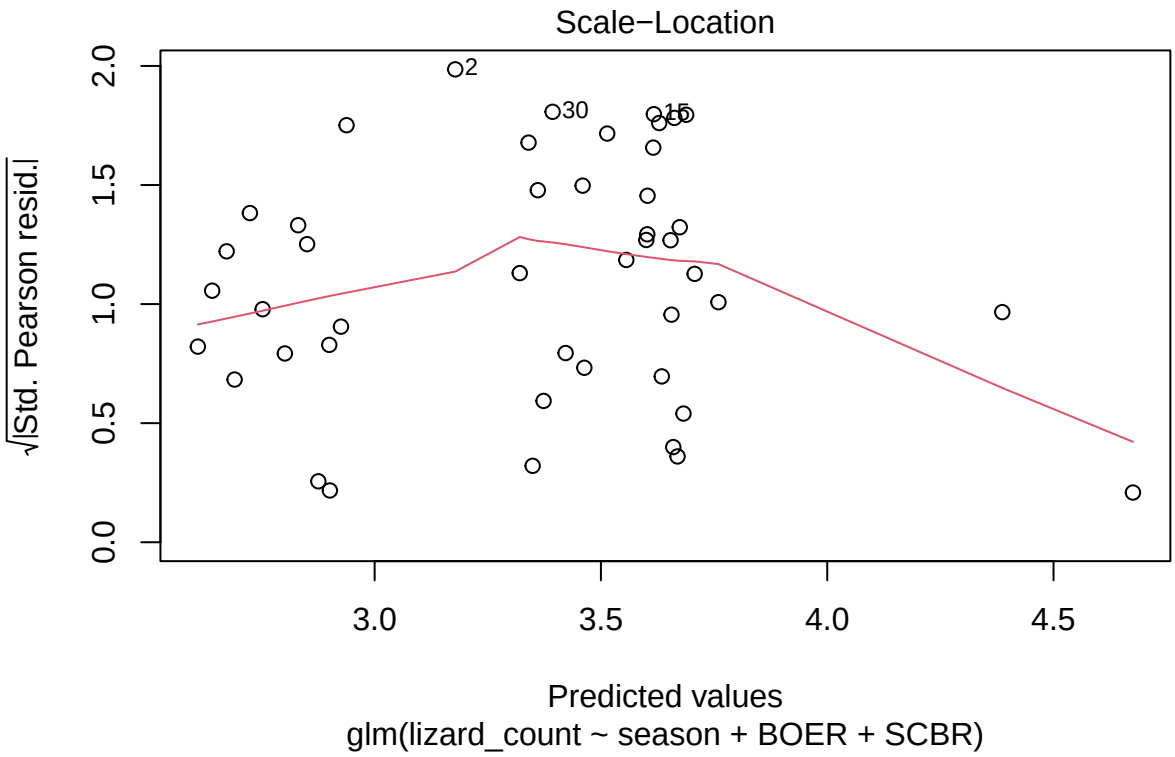
```
##
## Call:
## glm(formula = lizard_count ~ season + BOER + SCBR, family = "poisson",
##     data = npp_lizard3)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.672573   0.088936  41.295  < 2e-16 ***
## seasonS     -0.809719   0.071644 -11.302  < 2e-16 ***
## seasonW      0.001136   0.060477   0.019  0.98502
## BOER        -0.086618   0.029780  -2.909  0.00363 **
## SCBR         0.500638   0.040074  12.493  < 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: 462.06  on 44  degrees of freedom
## Residual deviance: 142.06  on 40  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 383.14
##
## Number of Fisher Scoring iterations: 4
```

```
plot(npp_lizard_mod3)
```

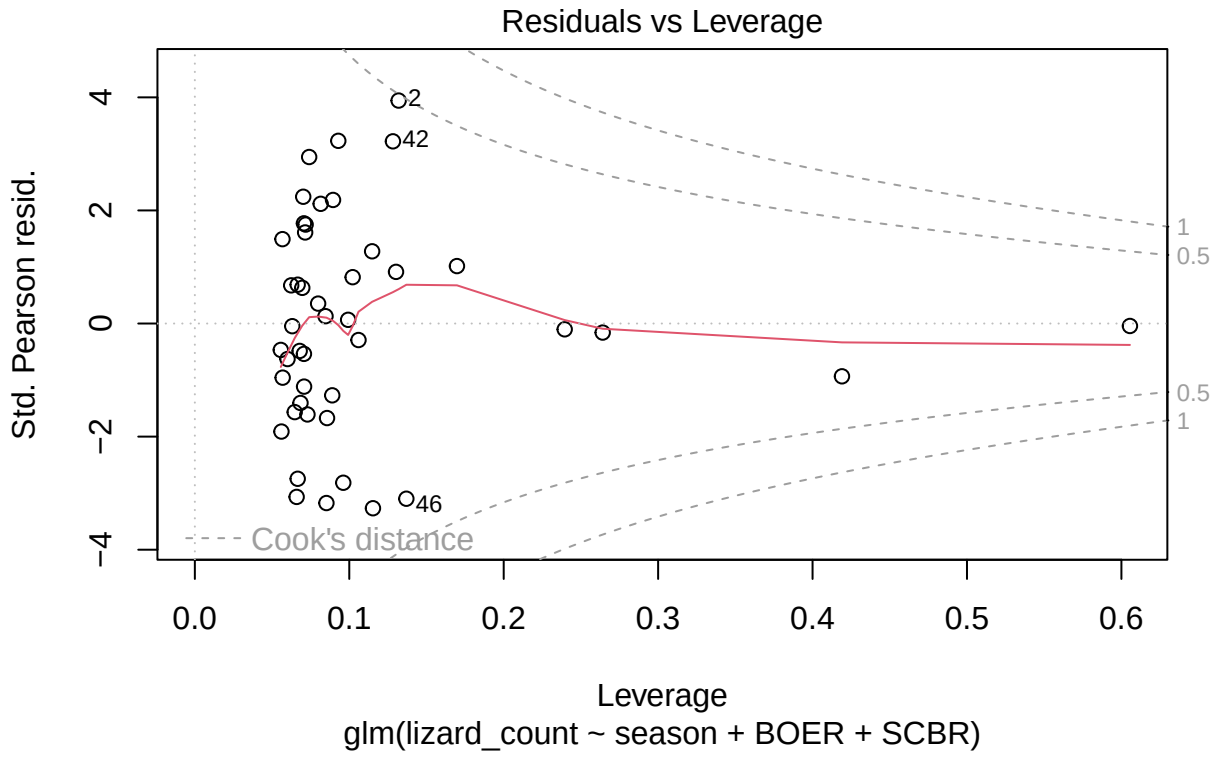



Q−Q Residuals
|Std. Deviance resid.|
0
1
2
3
4
0.0
0.5
1.0
1.5
2.0
2.5
30
2
7
Theoretical Quantiles
glm(lizard_count ~ season + BOER + SCBR)


Scale−Location
√|Std. Pearson resid.|
2.0
1.5
1.0
0.5
0.0
2
30
15
3.0
3.5
4.0
4.5
Predicted values
glm(lizard_count ~ season + BOER + SCBR)

Characteristic	log(IRR)	95% CI	p-value
season			
F	—	—	
S	-0.81	-0.95, -0.67	<0.001
W	0.00	-0.12, 0.12	>0.9
BOER	-0.09	-0.15, -0.03	0.004
SCBR	0.50	0.42, 0.58	<0.001

Abbreviations: CI = Confidence Interval, IRR = Incidence Rate Ratio




```
# Formate model results as gt table
tbl_regression(npp_lizard_mod3)
```

Step 5: Communicate Results

Results: The results of the poisson regression suggest that the summer season has a strong negative effect on lizard counts (B = -0.81, p < 0.0001) when compared to fall while winter has no strong effect (B = 0.001, p = 0.99). Black grama grass had a strong negative influence on lizard counts (B = -0.087, p = 0.004) while burrograss had a strong positive influence (B

= 0.50, p < 0.001). Note, the coefficents are for estimation of log-transformed counts and 4 outlier points we removed because they fell outside Cook's distance.

(3) See Github - nholsclaw