Lab Assignment 9

Summer Heschong

2025-03-25

(1) Logistic Regression

```
#load packages
library(here)
```

```
## here() starts at /Users/summerheschong/stats_spring25
```

```
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)
library(DHARMa)
```

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

```
library(gtsummary)
library(broom)

#load data
lizards <- read.csv(here('Data/Raw/jrn_lizard.csv'))

#filter dataset for side-blotched lizards
lizards <- lizards %>%
  filter(spp == 'UTST') %>%
  na.omit
```

Step 1 - Define research question

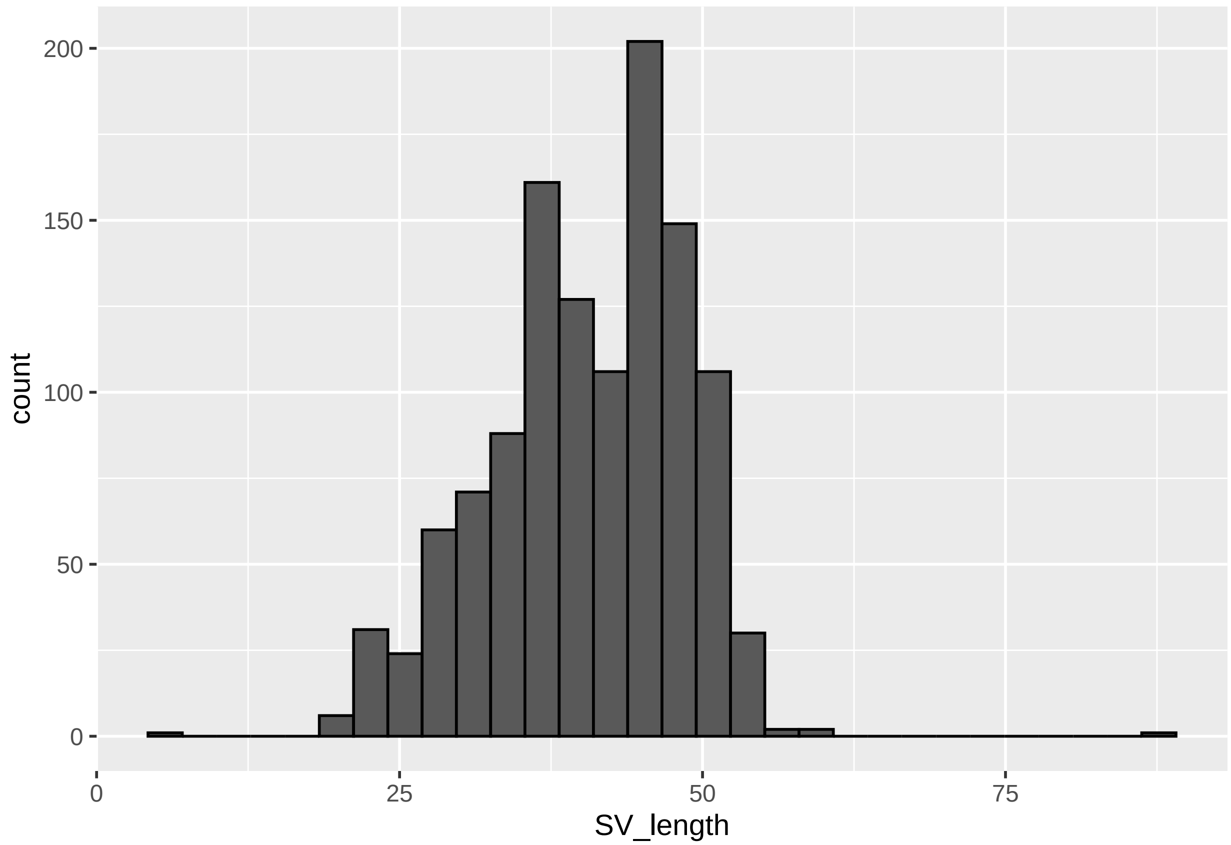
Do snout-to-vent length, sex, and vegetation zone at time of capture significantly predict if a lizard tail is recorded as whole?

Step 2 - Examine data

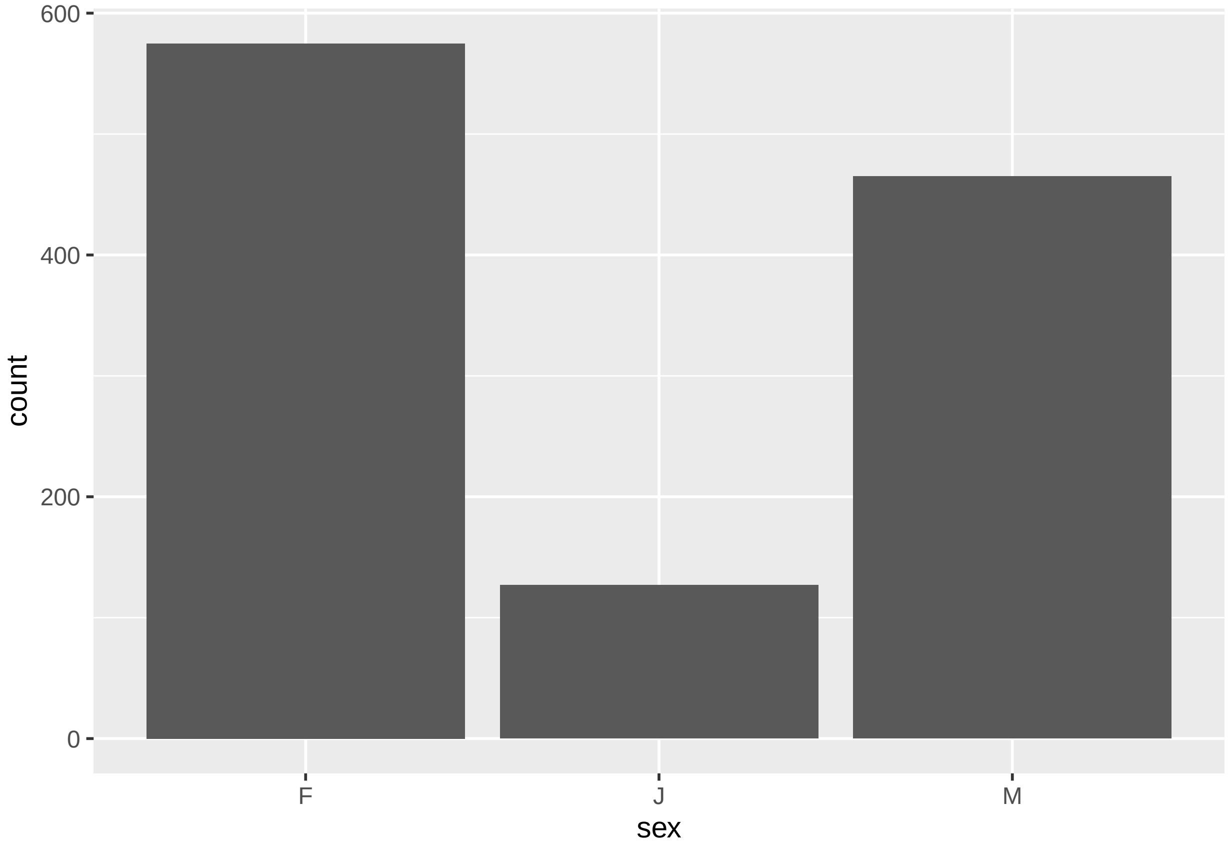
a. Display distributions or raw counts of data

```
#create histograms displaying data

#snout-to-vent length
ggplot(lizards, aes(x = SV_length)) +
geom_histogram(color = 'black')
```

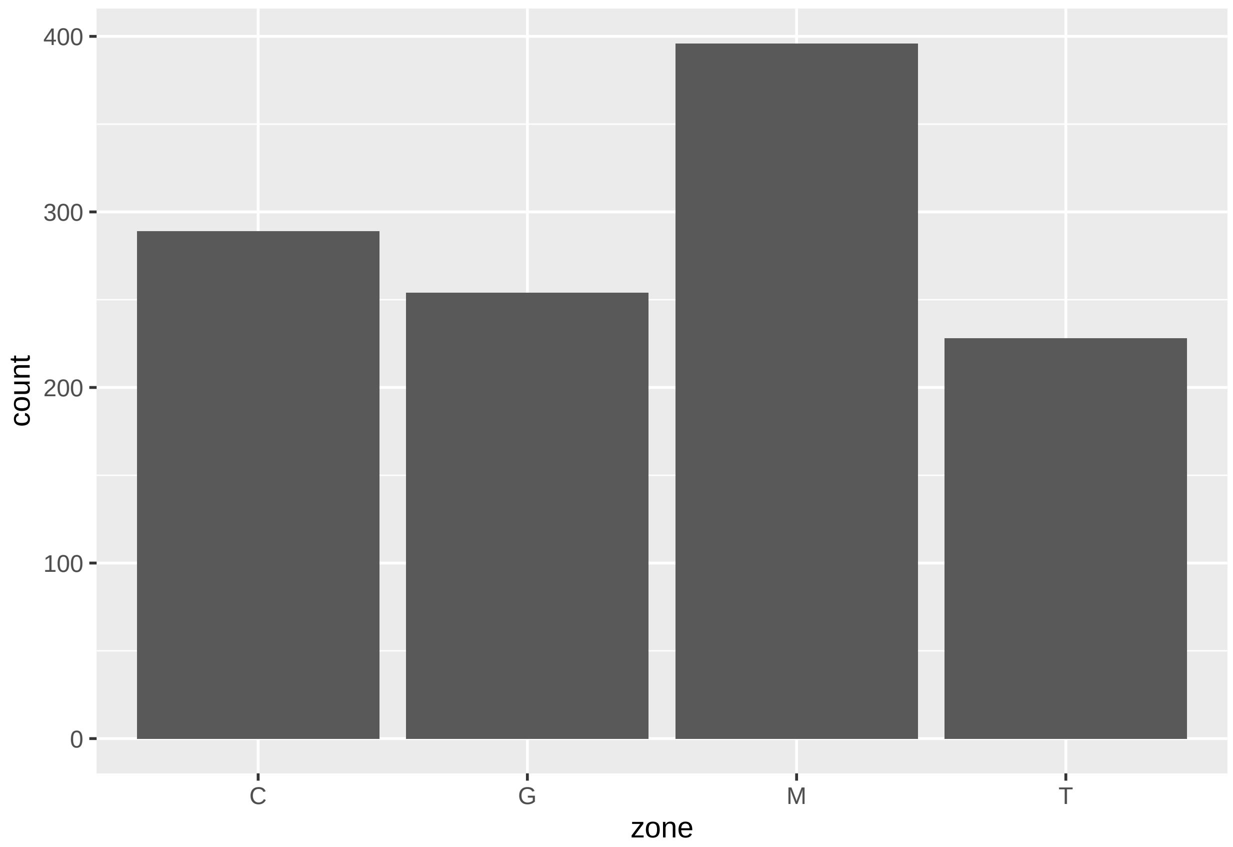
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



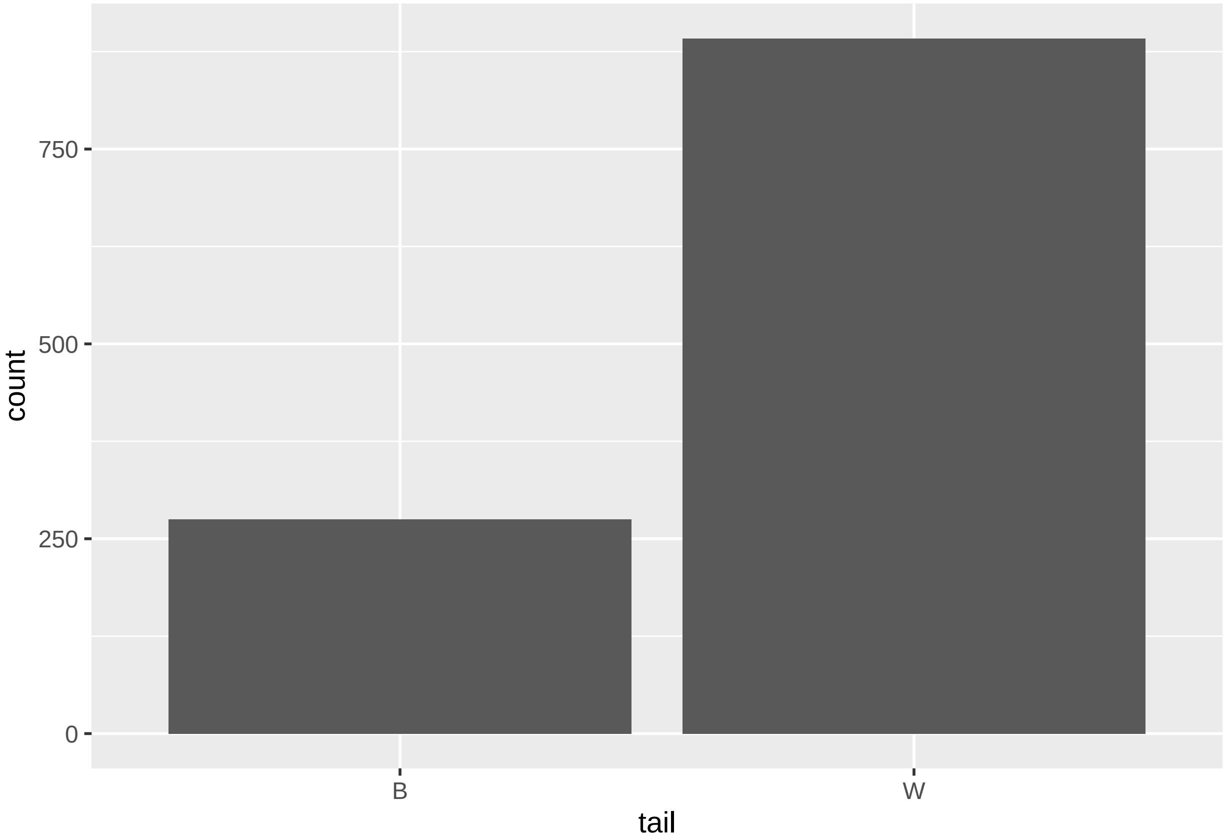
```
#sex
ggplot(lizards, aes(x = sex)) +
geom_bar ()
```



```
#vegetation zone
ggplot(lizards, aes(x = zone)) +
geom_bar ()
```

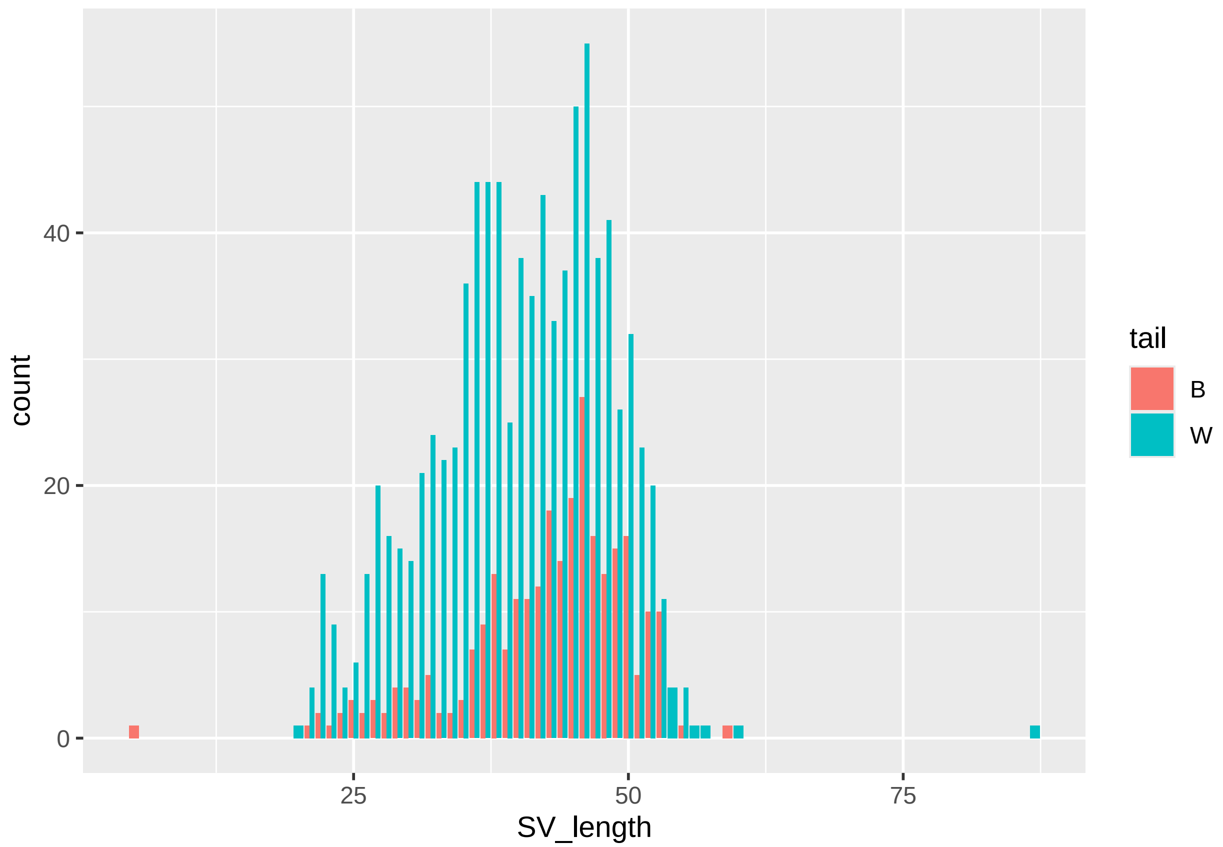


```
#lizard tail status
ggplot(lizards, aes(x = tail)) +
geom_bar ()
```

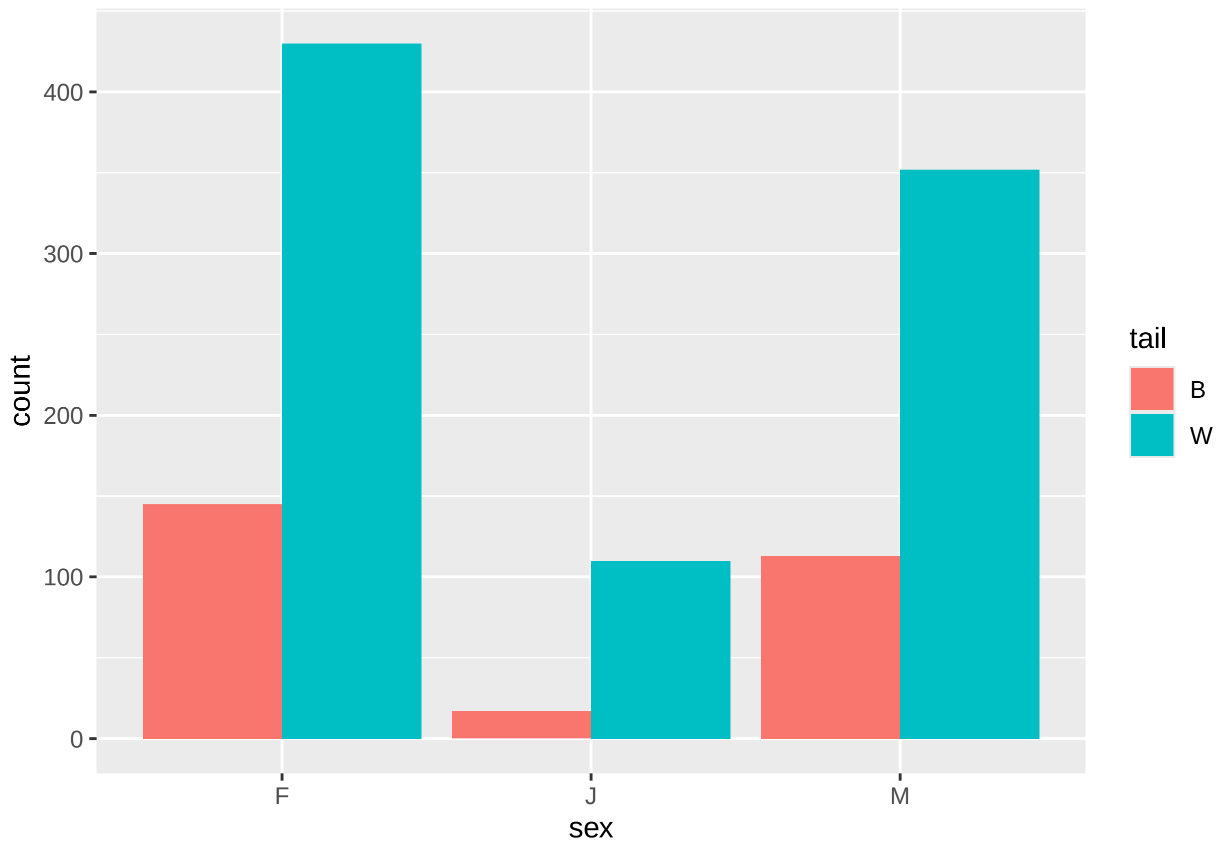


b. Display relationships between predictor and outcome variables

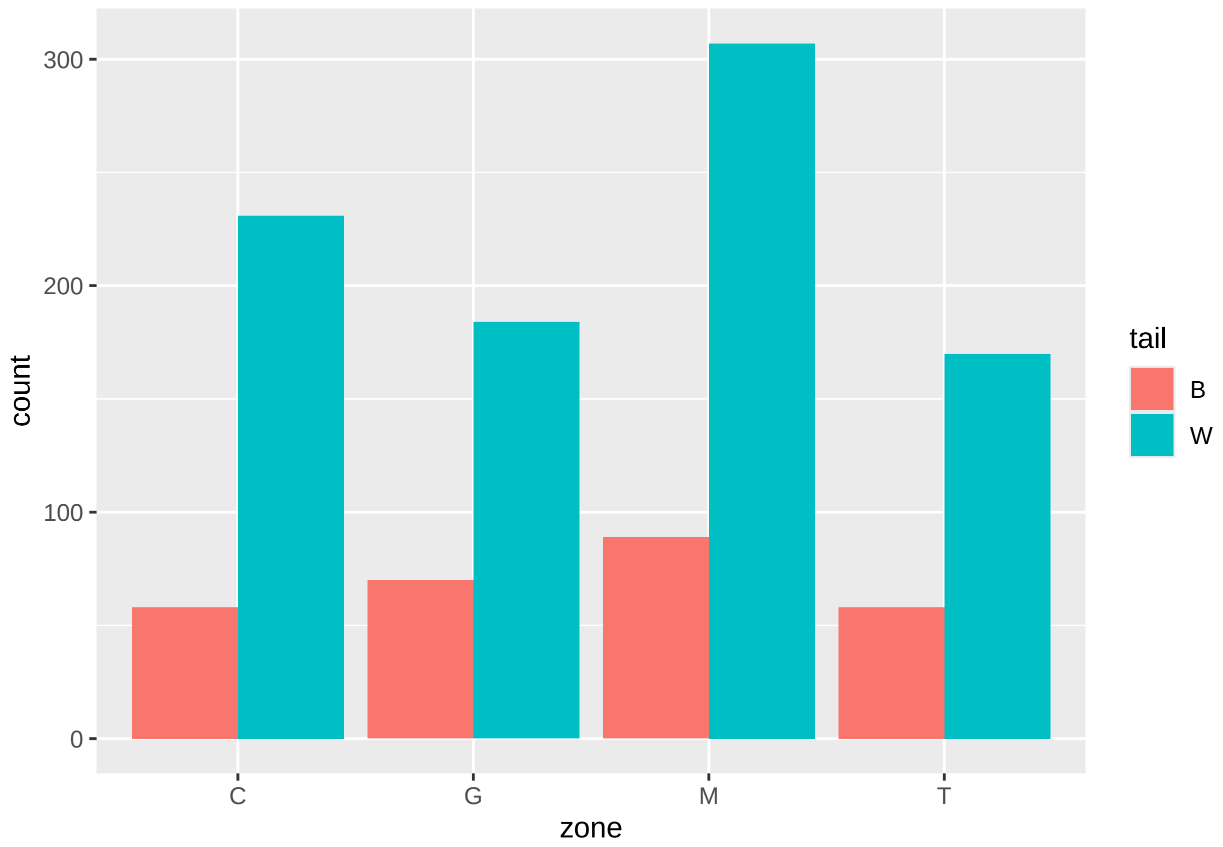
```
#create bar plots

#snout-to-vent length
ggplot(lizards, aes(x = SV_length, fill = tail)) +
geom_bar (position = 'dodge')
```



```
#sex
ggplot(lizards, aes(x = sex, fill = tail)) +
geom_bar (position = 'dodge')
```



```
#zone
ggplot(lizards, aes(x = zone, fill = tail)) +
geom_bar (position = 'dodge')
```



Step 3 - Fit regression model

```
# change characters to factors
lizards$sex <- factor(lizards$sex, levels = c('J', 'F', 'M'))

lizards$zone <- factor(lizards$zone, levels = c('G', 'C', 'M', 'T'))

lizards$tail <- factor(lizards$tail, levels = c('B', 'W'))

#fit regression model
tail_mod <- glm(tail ~ SV_length + sex + zone,
                data = lizards,
                family = 'binomial')
```

Step 4 - Evaluate model diagnostics

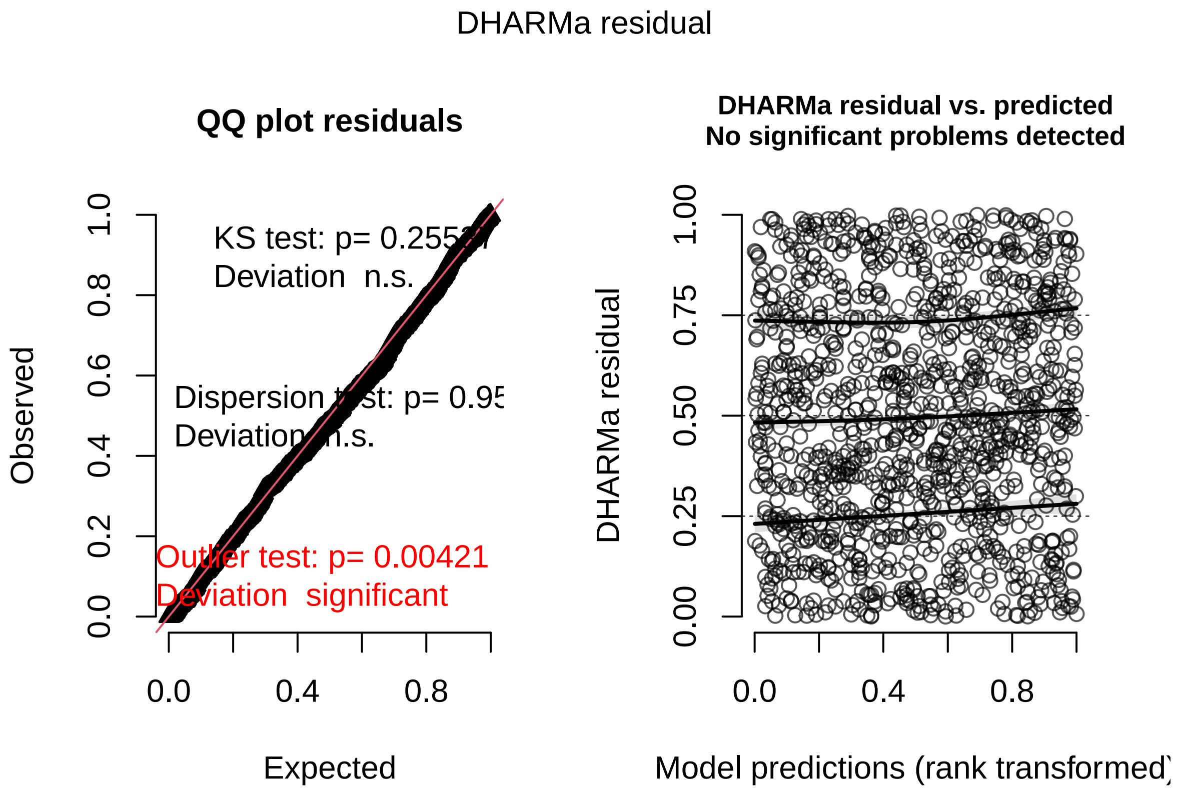
```
#examine model output
summary(tail_mod)
```

```
##
## Call:
```

```
## glm(formula = tail ~ SV_length + sex + zone, family = "binomial",
##     data = lizards)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.85198    0.43630   6.537 6.29e-11 ***
## SV_length   -0.04075    0.01083  -3.762 0.000169 ***
## sexF        -0.33473    0.30241  -1.107 0.268356
## sexM        -0.20694    0.31869  -0.649 0.516114
## zoneC        0.50862    0.20800   2.445 0.014473 *
## zoneM        0.27847    0.18718   1.488 0.136834
## zoneT        0.25799    0.21165   1.219 0.222866
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1274.4  on 1166  degrees of freedom
## Residual deviance: 1245.7  on 1160  degrees of freedom
## AIC: 1259.7
##
## Number of Fisher Scoring iterations: 4
```

```r
#examine modified residuals plot
simulateResiduals(tail_mod) %>% plot()
```

```
## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```



Step 5 - Interpret model and communicate results

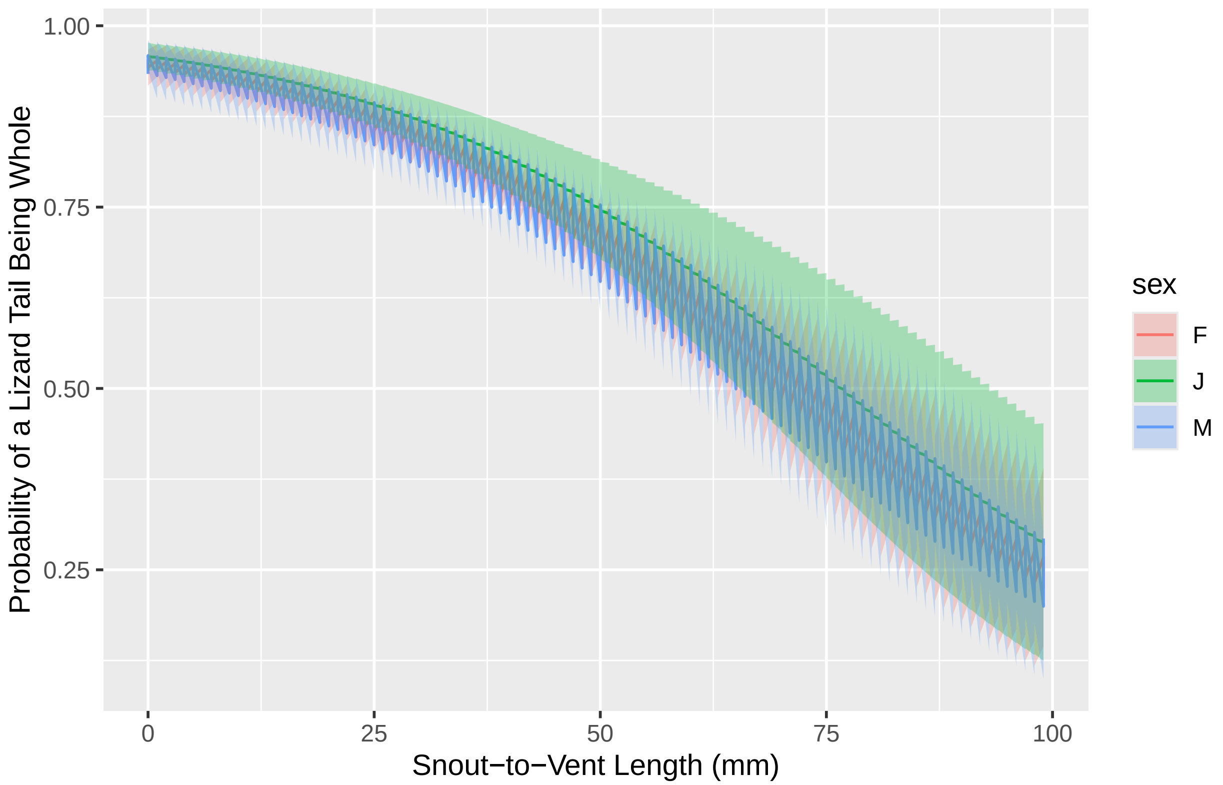
a. Report results

The results of a logistic regression suggest that the log-odds of a lizard tail being whole are greater with shorter snout-to-vent length (B = -0.04, p < 0.001). There was no significant relationship between the log-odds of a lizard tail being whole and the sex of the lizard relative to juvenile lizards (male: B = -0.21, p < 0.52; female: B = -0.33, p < 0.27). There was a significant increase in the log-odds of a lizard tail being whole when the lizard is captured in creosotebush shrubland (B = 0.51, p < 0.01) relative to grama grassland. However, lizards caught in mesquite duneland and tarbush shrubland did not have significantly different log-odds of a lizard tail being whole relative to those caught in grama grassland (zoneM: B = 0.28, p < 0.14; zoneT: B = 0.26, p < 0.22).

b. Provide 2 predictive figures

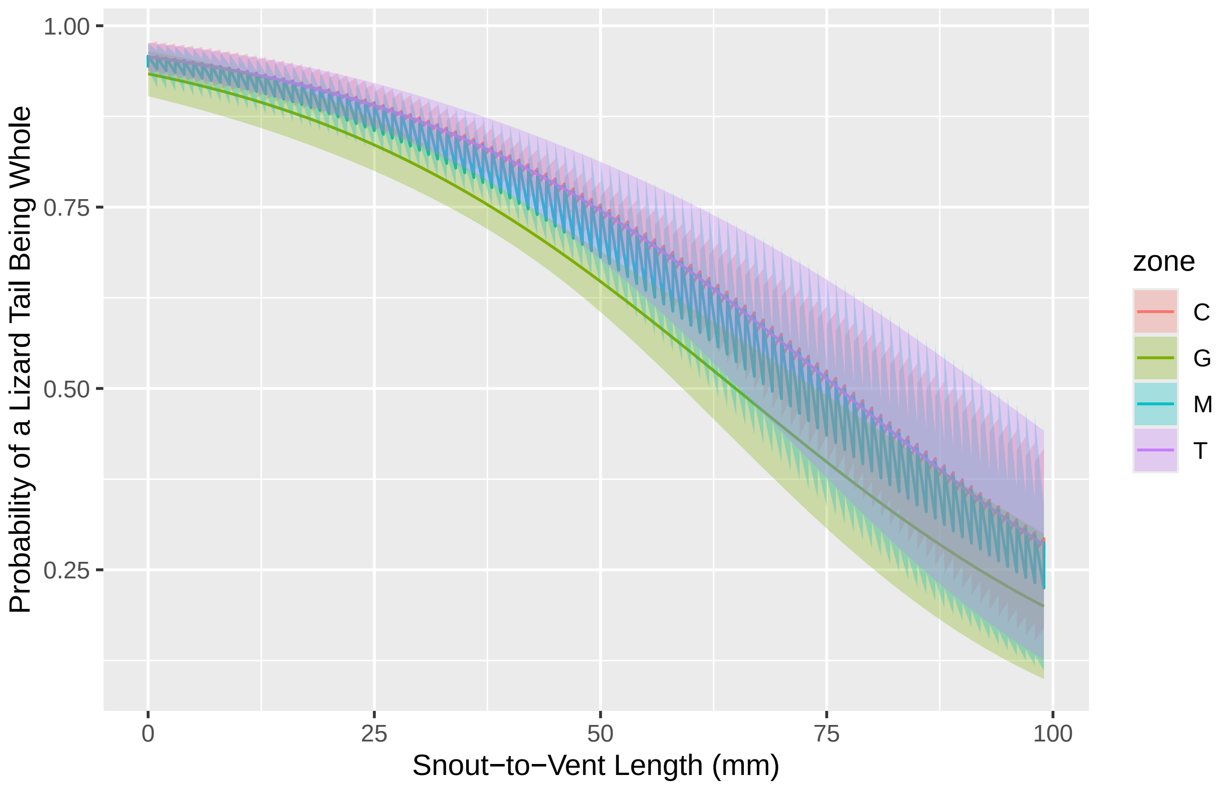
```r
# simulate snout-to-vent length data
SV_length_vector <- rep(seq(from = 0, to = 99), 12)

#simulate sex data
sex_vector <- c(rep('M', 400),
                rep('F', 400),
                rep('J', 400))
```

```r
#simulate zone data
zone_vector <- c(rep('G', 300),
                 rep('C', 300),
                 rep('M', 300),
                 rep('T', 300))

#join data
pred_data <- data.frame(SV_length_vector, sex_vector, zone_vector)
colnames(pred_data) <- c('SV_length', 'sex', 'zone')

#Use original model to predict outcomes
prediction <- predict(tail_mod,
                      newdata = pred_data,
                      type = 'respons',
                      se.fit = TRUE)
#pull out predictions
prob_data  <- data.frame(pred_data,
                        prediction$fit,
                        prediction$se.fit)
#rename columns
colnames(prob_data) <- c('SV_length', 'sex', 'zone', 'probability', 'se')

#Graph probabilities of lizard tail being whole
ggplot(prob_data, 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 Lizard Tail Being Whole',
  caption = 'Predictive figure showing change in probability of a lizard tail being whole \n as lizard
```



Predictive figure showing change in probability of a lizard tail being whole
as lizard length increases, with lines colored by sex.

```r
ggplot(prob_data, 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 Lizard Tail Being Whole',
   caption = 'Predictive figure showing change in probability of a lizard tail being whole \n as lizard
```



Predictive figure showing change in probability of a lizard tail being whole as lizard length increases, with lines colored by vegetation zone.

(2) Poisson Regression

Step 1 - Define research question

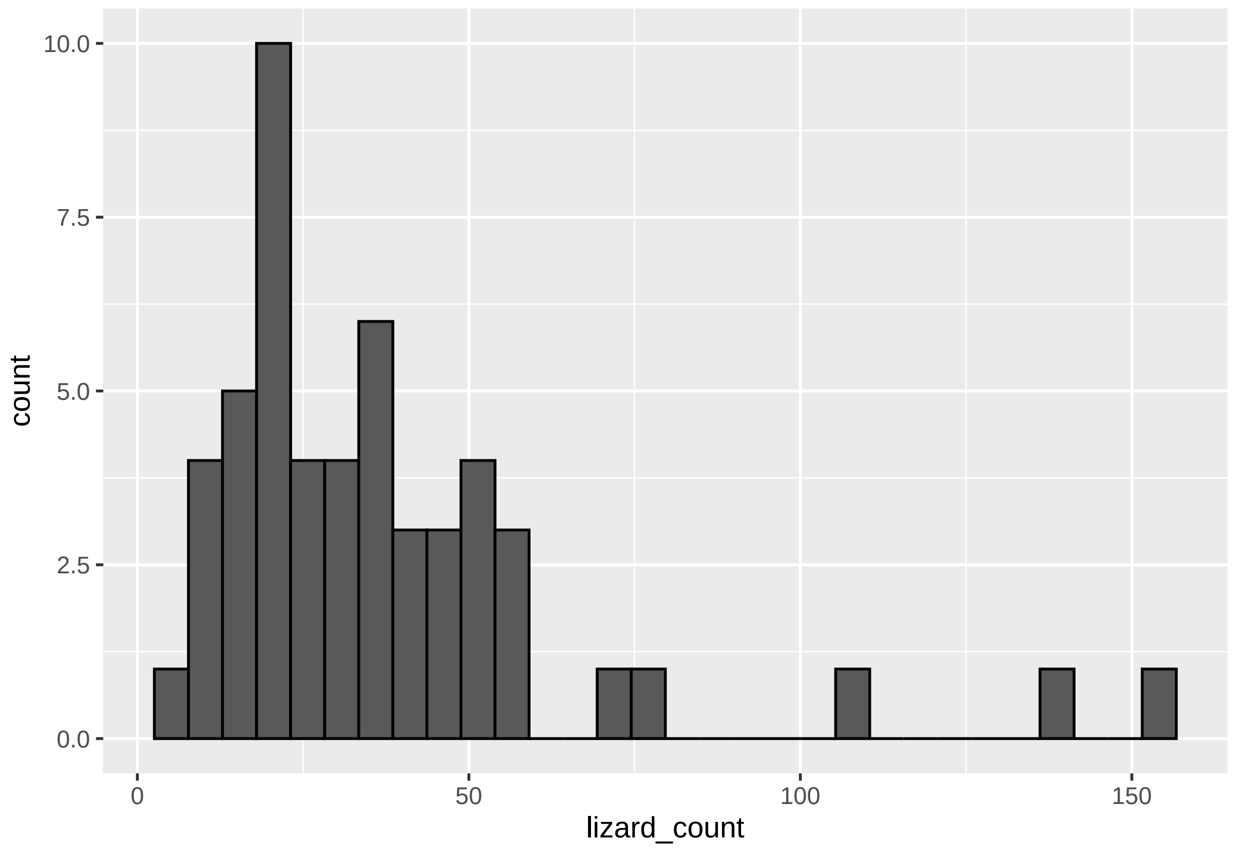
Do season and plant species percent cover significantly predict lizard counts?

```
lizard_counts <- read.csv(here('Data/Raw/jrn_lizard_npp.csv'))
```

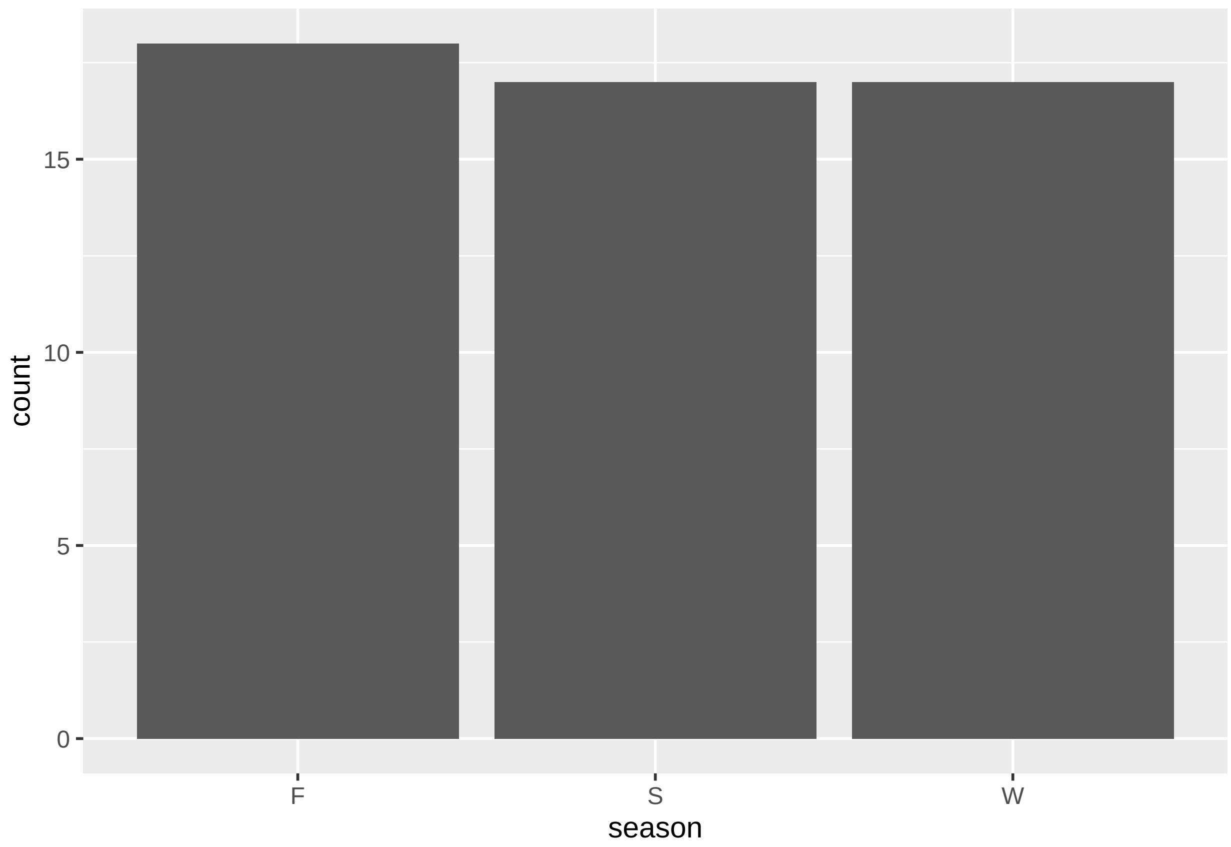
Step 2 - Examine data and possible correlations

```
#create histograms
ggplot(lizard_counts, aes(x = lizard_count)) +
geom_histogram(color = 'black')
```

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

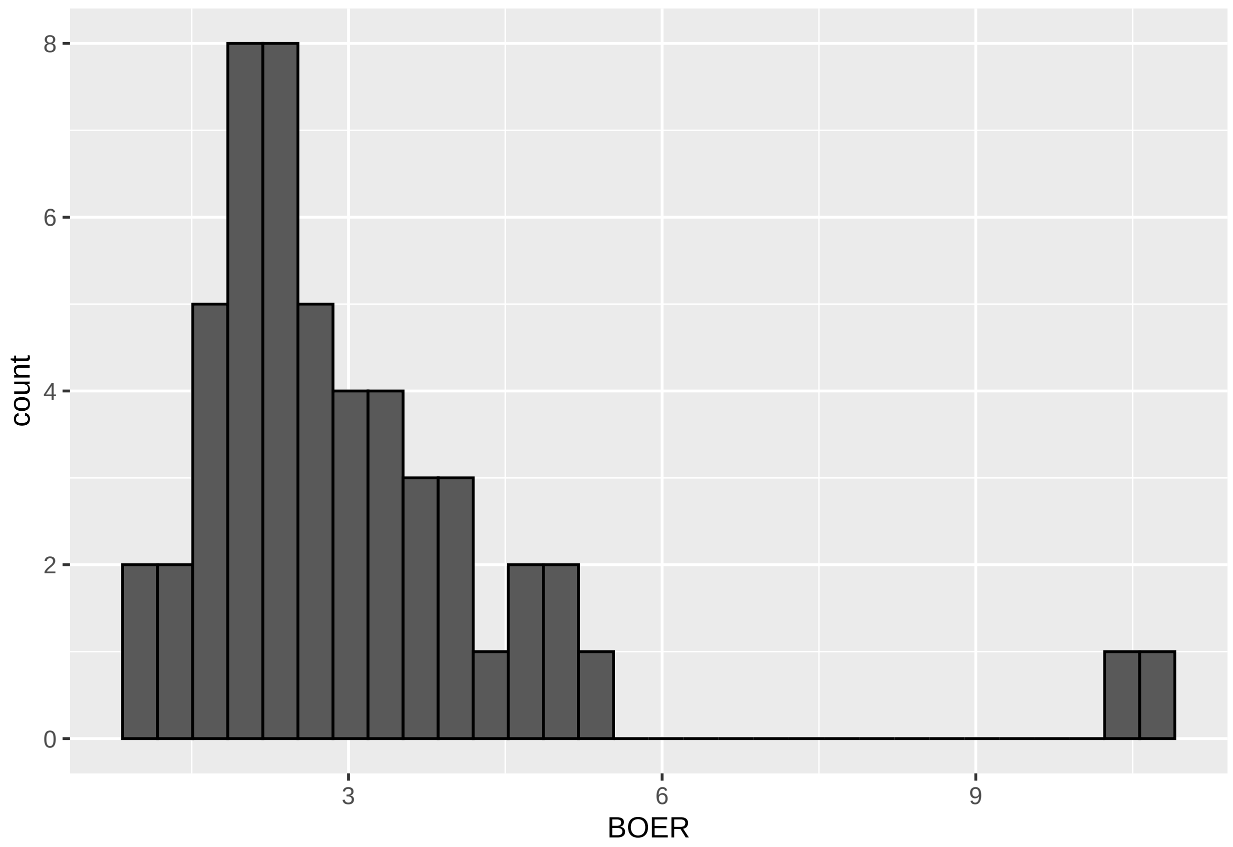


```
ggplot(lizard_counts, aes(x = season)) +
geom_bar()
```



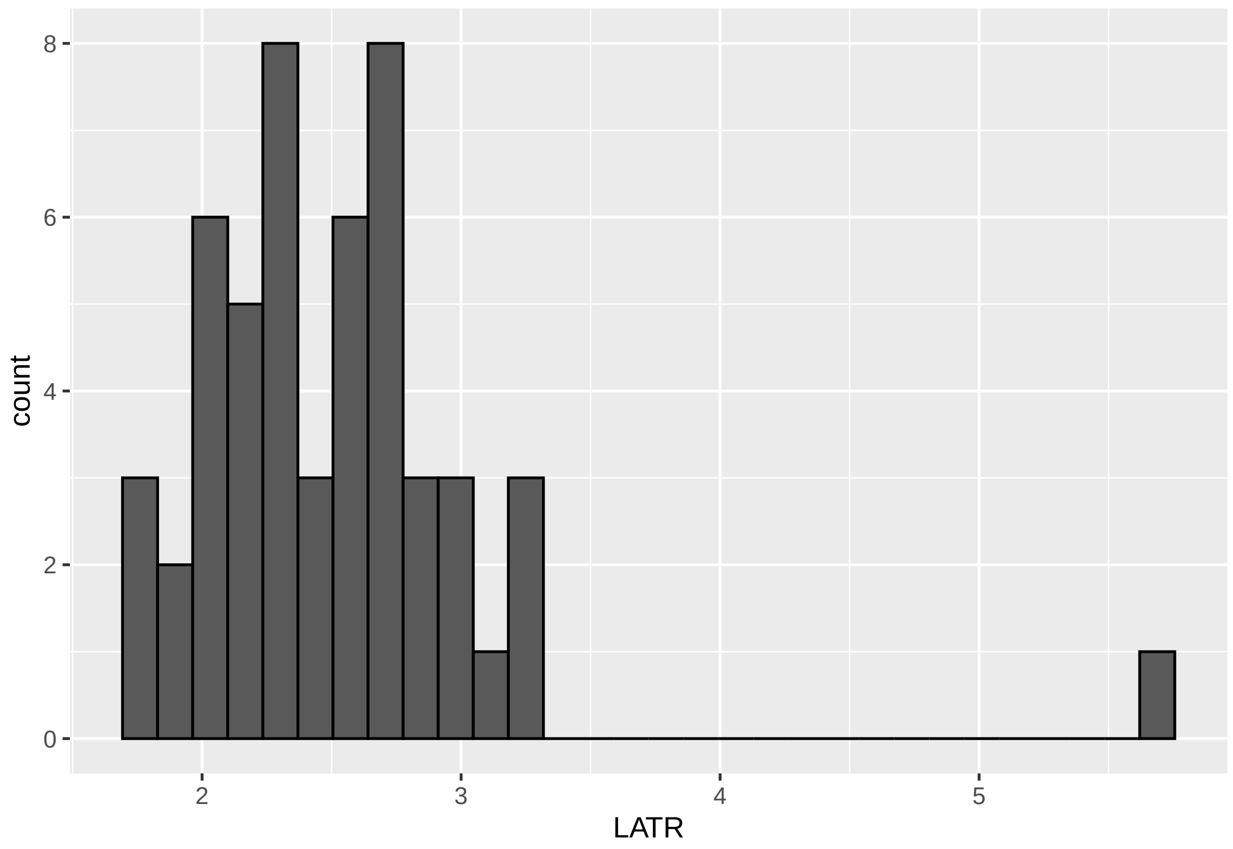
```
ggplot(lizard_counts, aes(x = BOER)) +
geom_histogram(color = 'black')
```

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



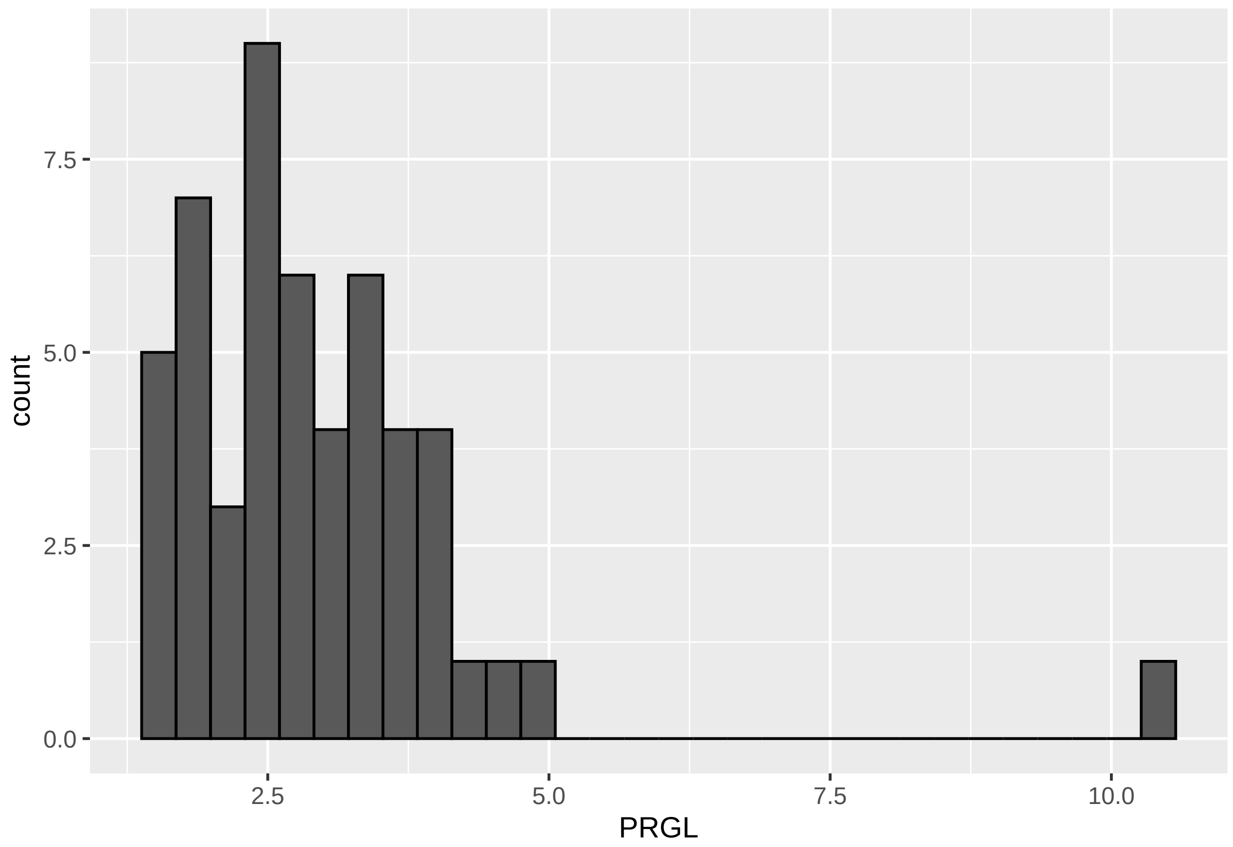
```
ggplot(lizard_counts, aes(x = LATR)) +
geom_histogram(color = 'black')
```

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



```r
ggplot(lizard_counts, aes(x = PRGL)) +
geom_histogram(color = 'black')
```

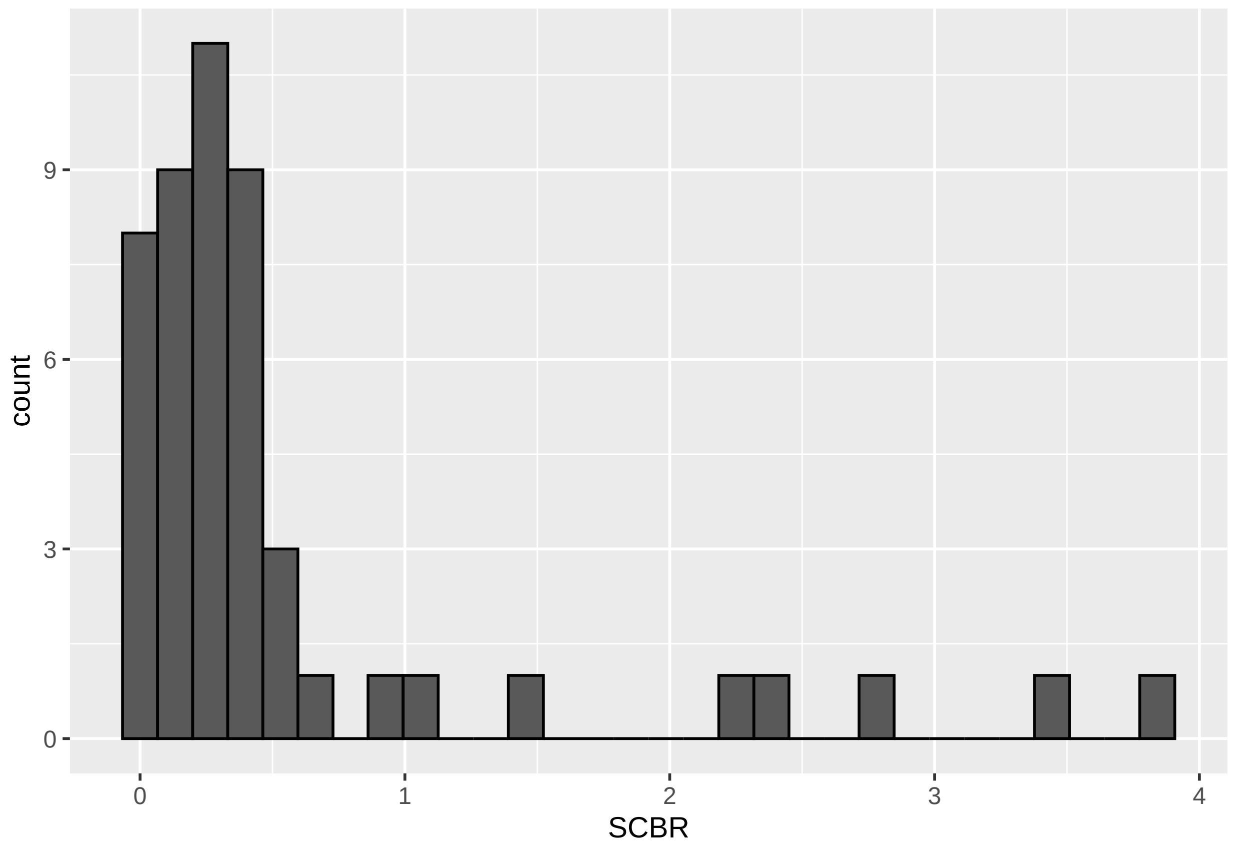
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



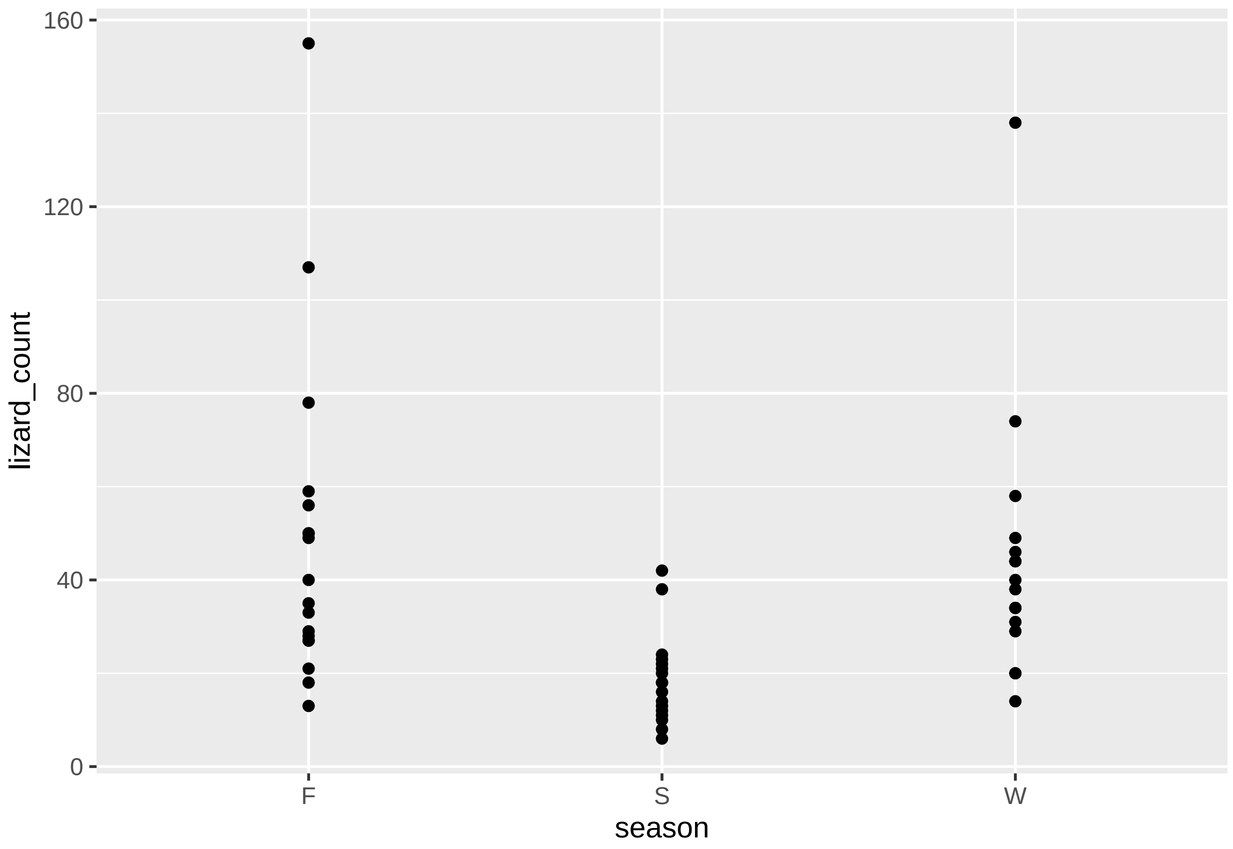
```
ggplot(lizard_counts, aes(x = SCBR)) +
geom_histogram(color = 'black')
```

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

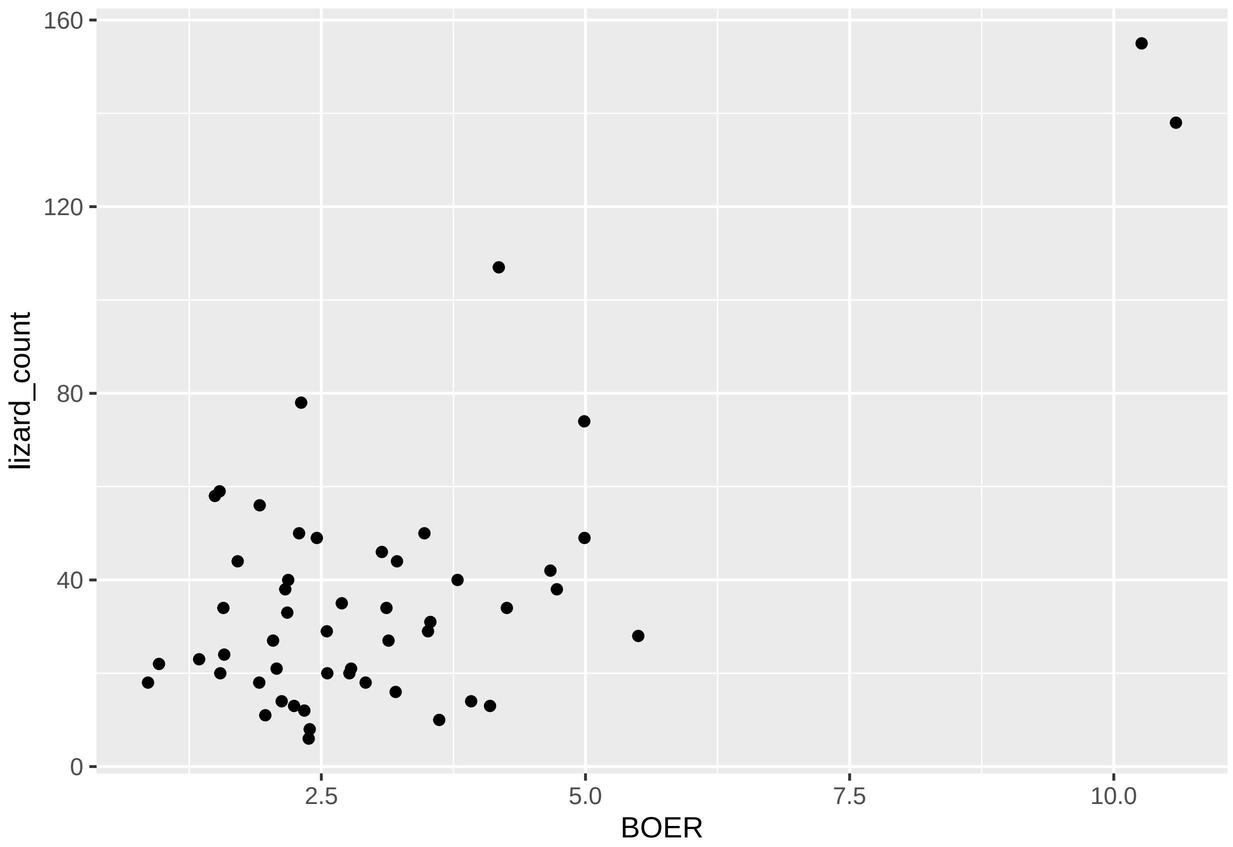
```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```



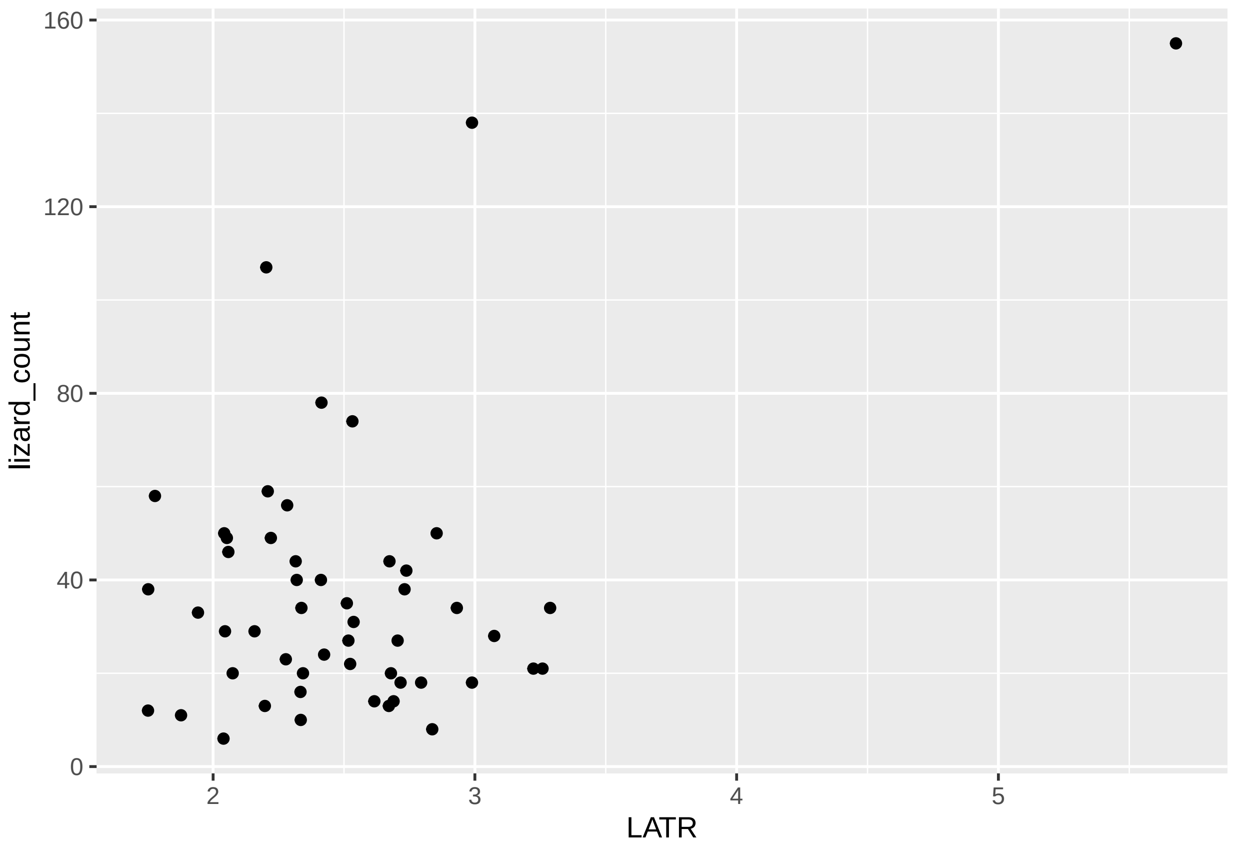
```
#Create scatter plots
ggplot(lizard_counts, aes(x = season, y = lizard_count)) +
geom_point(color = 'black')
```



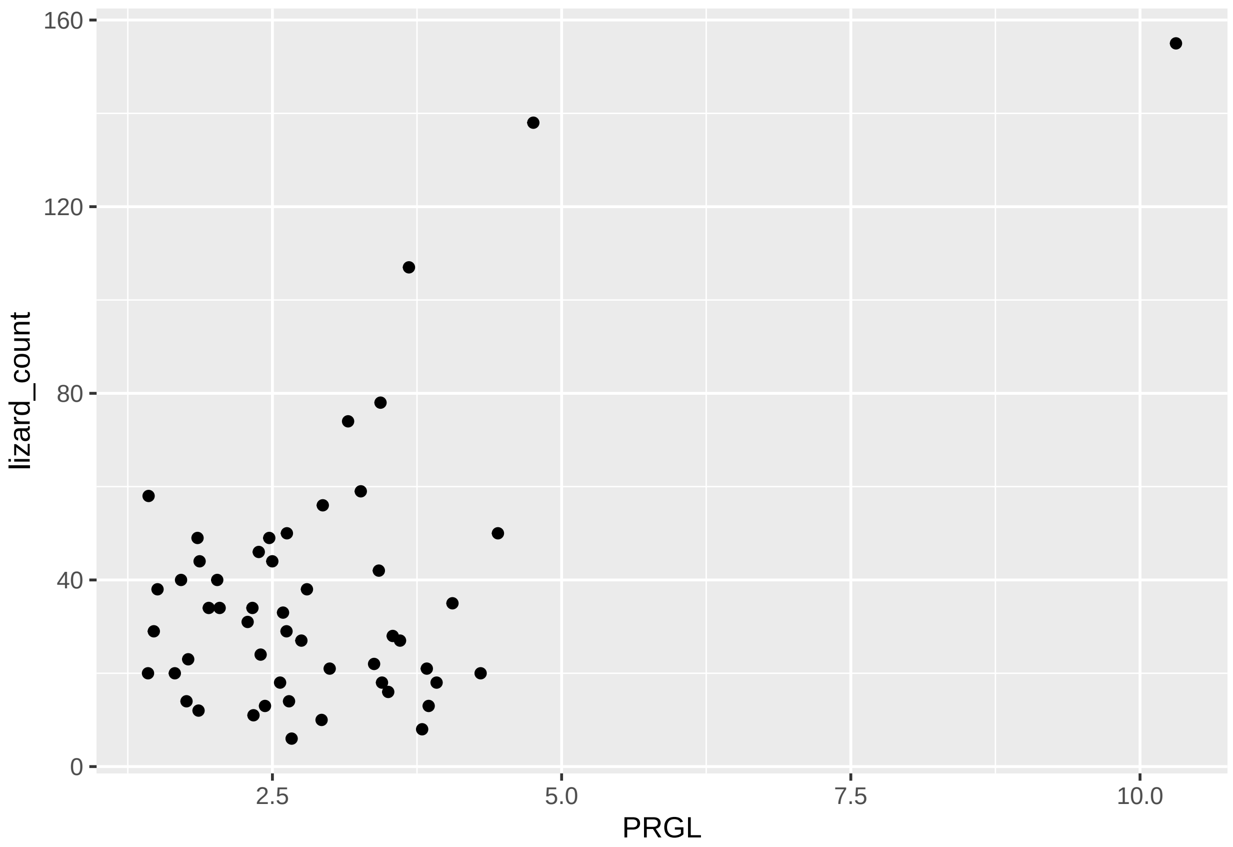
```
ggplot(lizard_counts, aes(x = BOER, y = lizard_count)) +
geom_point(color = 'black')
```



```
ggplot(lizard_counts, aes(x = LATR, y = lizard_count)) +
geom_point(color = 'black')
```

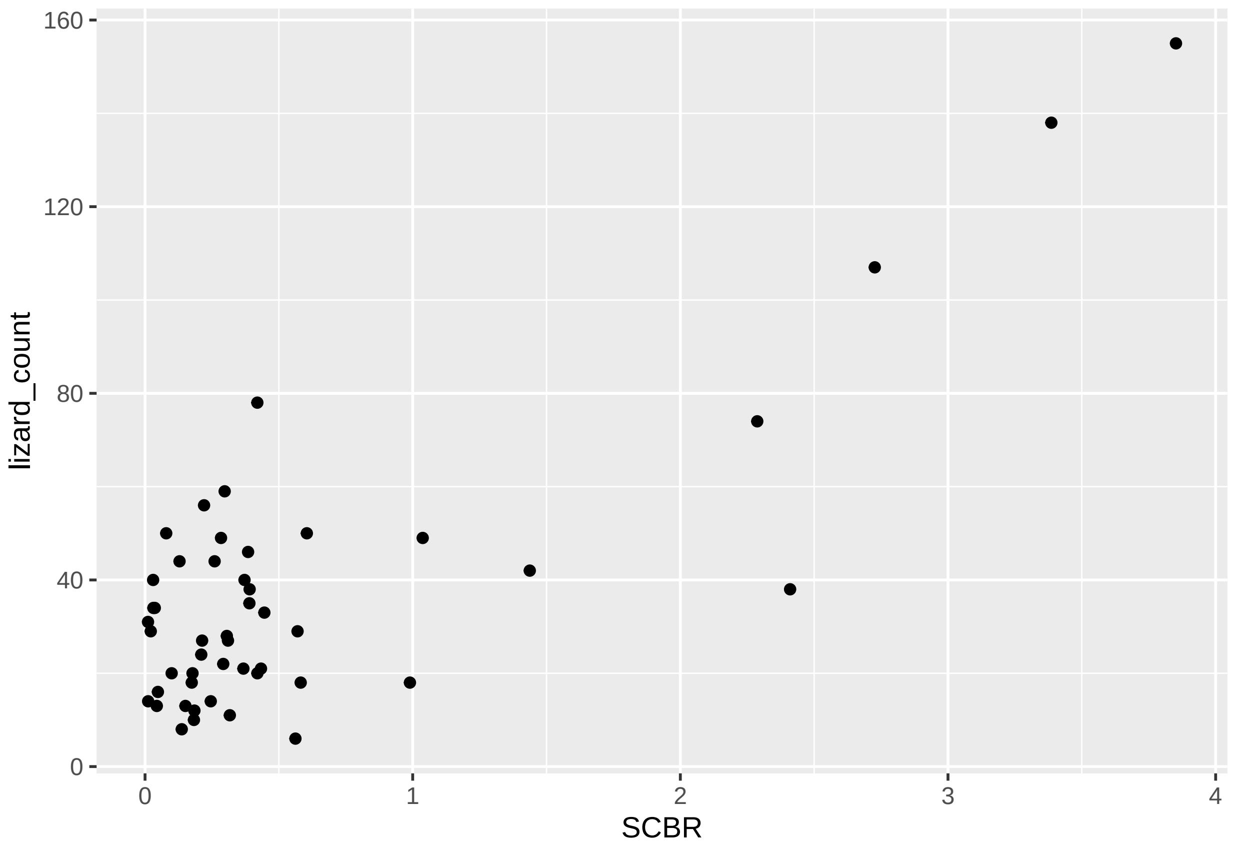


```r
ggplot(lizard_counts, aes(x = PRGL, y = lizard_count)) +
geom_point(color = 'black')
```



```r
ggplot(lizard_counts, aes(x = SCBR, y = lizard_count)) +
geom_point(color = 'black')
```

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



Step 3 - Fit regression model

```
#transform characters to factors
lizard_counts$season <- factor(lizard_counts$season,
                                 levels = c('W', 'S', 'F'))

#fit poisson regression model
lizard_mod <- glm(lizard_count ~ season + BOER + LATR + PRGL,
                  data = lizard_counts,
                  family = 'poisson')
```

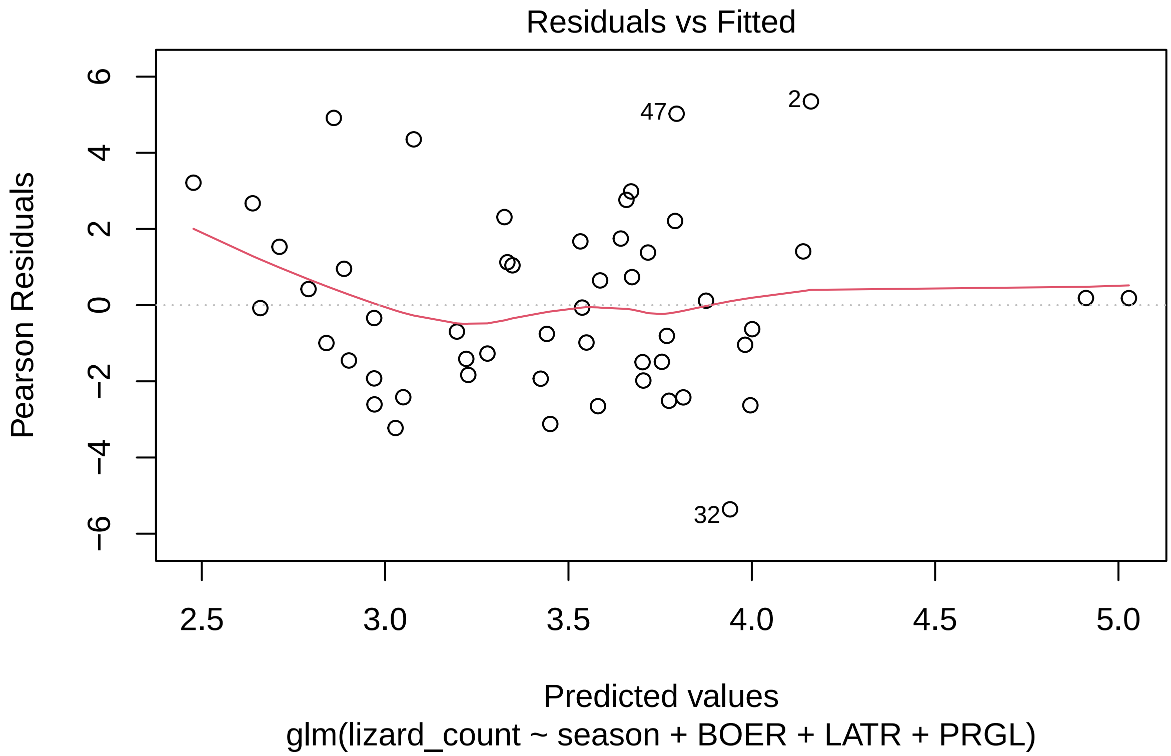
Step 4 - Evaluate model diagnostics

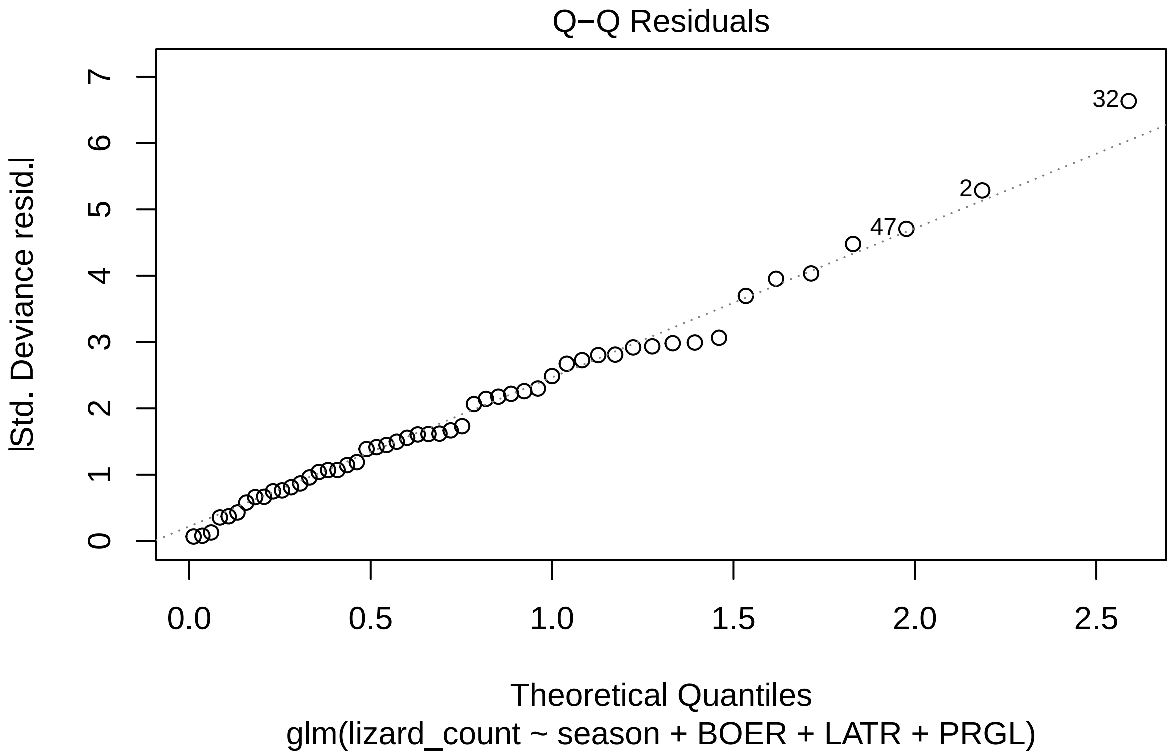
```
#examine model output
summary(lizard_mod)
```

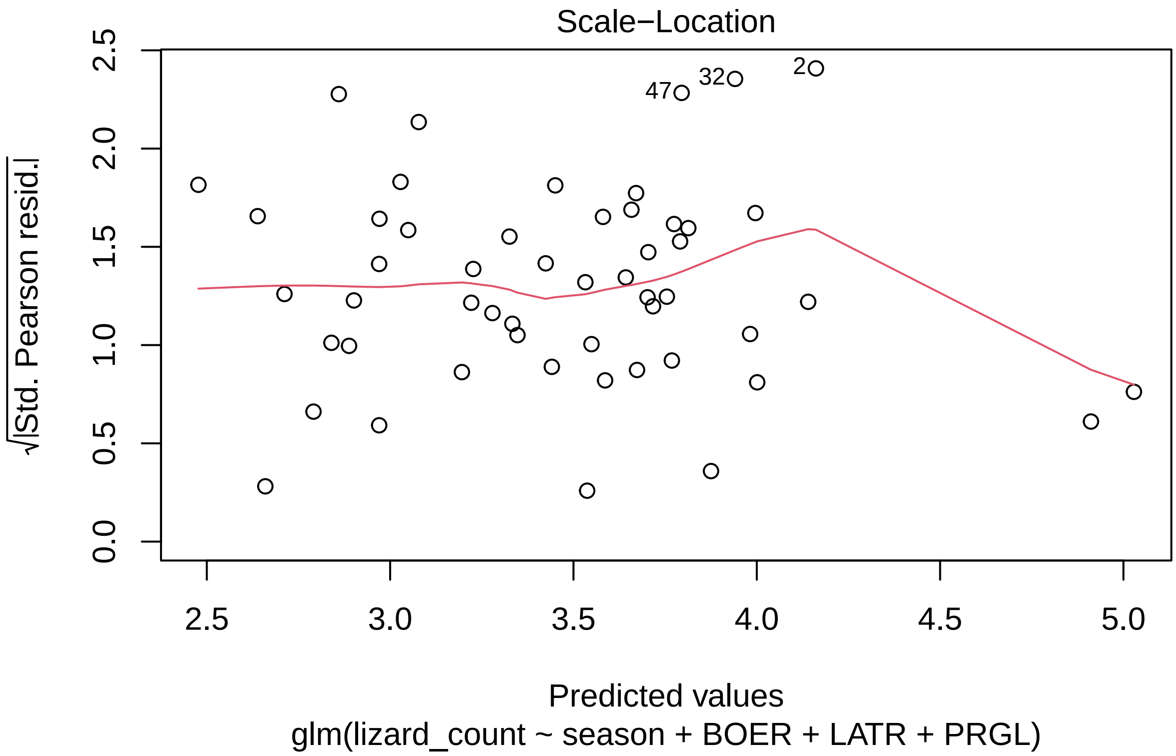
```
##
## Call:
## glm(formula = lizard_count ~ season + BOER + LATR + PRGL, family = "poisson",
##     data = lizard_counts)
##
```

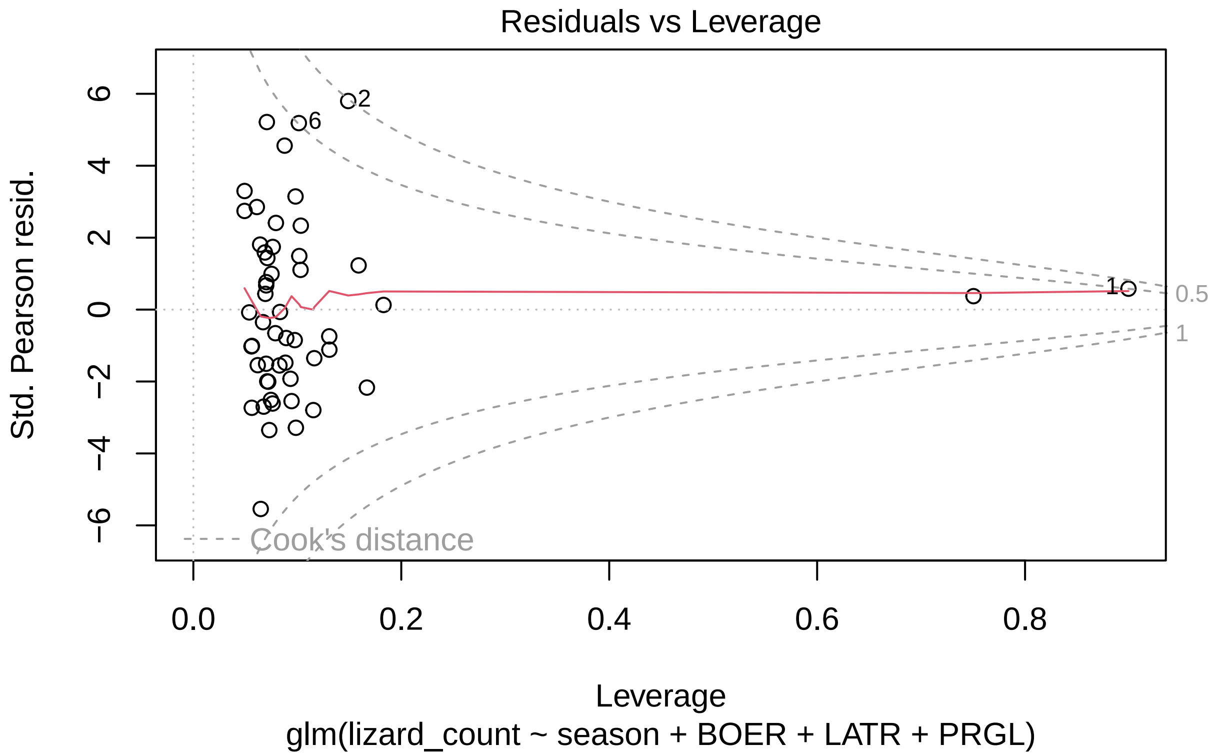
```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.07539    0.11690  34.861  < 2e-16 ***
## seasonS     -1.01184    0.08062 -12.551  < 2e-16 ***
## seasonF     -0.29924    0.07087  -4.222 2.42e-05 ***
## BOER         0.08139    0.01560   5.219 1.80e-07 ***
## LATR        -0.58951    0.07251  -8.130 4.29e-16 ***
## PRGL         0.36512    0.04009   9.107  < 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: 927.91  on 51  degrees of freedom
## Residual deviance: 278.16  on 46  degrees of freedom
## AIC: 562.07
##
## Number of Fisher Scoring iterations: 4
```

```
#examine model residuals
plot(lizard_mod)
```




Q−Q Residuals
|Std. Deviance resid.|
0
1
2
3
4
5
6
7
0.0
0.5
1.0
1.5
2.0
2.5
Theoretical Quantiles
glm(lizard_count ~ season + BOER + LATR + PRGL)
32
2
47


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
2.5
2.5
3.0
3.5
4.0
4.5
5.0
47
32
2
Predicted values
glm(lizard_count ~ season + BOER + LATR + PRGL)



Step 5 - Interpret the model and communicate the results

The results of a poisson regression suggest that summer (B = -1.01, p < 0.001) and fall (B = -0.30, p <0.001) have significantly lower side-blotched lizard counts relative to winter. Also lizard counts were found to be significantly greater as percent cover of black grama grass (B = 0.08, p < 0.001) and honey mesquite (B = 0.37, p < 0.001) increase, and percent cover of creosote bush decreases (B = -0.59, p < 0.001).

(3) GitHub customization

https://github.com/srheschong