Lab 14

Nitasha Fazal, Meghan Robinson, Zoe Zarvatski

Rules

In groups of 2 or 3, complete the following.

Modeling Snow Presence, version 2

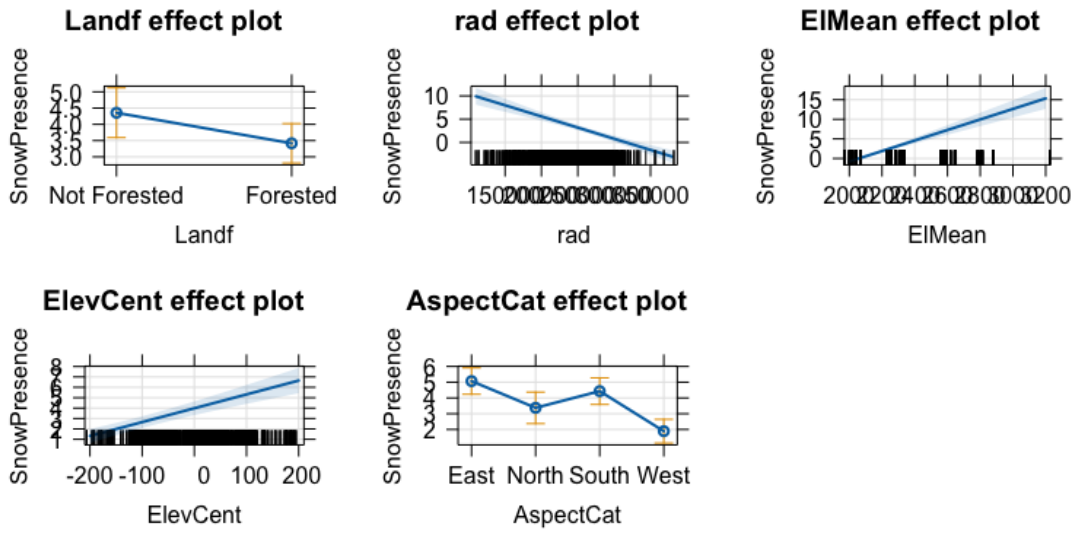
We will continue to use the data set from Wetlaufer, Hendrikx, and Marshall (2016) that explored the relationship between snow density (kg/m^3) or snow depth (`snow`, mm) with a suite of predictor variables. To be able to measure characteristics of snow, they needed to find snow in the locations they were sampling, so the focus in this lab will be on the snow presence or absence at each location (`SnowPresence`). We will be interested in using `elev` (Elevation, m), `Land` (forest cover with 0 = unforested and 10 = forested), `rad` (Potential Solar radiation, Wh/m^2), `curvature` (see https://blogs.esri.com/esri/arcgis/2010/10/27/understanding-curvature-rasters/ for a description), `aspect` (orientation of slope in degrees (0 to 360)), and `angle` (angle of slope in degrees with 0 being flat) as fixed effect predictors. Also pay attention to the `strata` variable (read its definition in the paper) and the role that played in the data collection and should in the analysis.

- Wetlaufer, K., Hendrikx, J., and L. Marshall (2016) Spatial Heterogeneity of Snow Density and Its Influence on Snow Water Equivalence Estimates in a Large Mountainous Basin. *Hydrology*, 3(1):3, doi:10.3390/hydrology3010003. Available at http://www.mdpi.com/2306-5338/3/1/3/htm and on D2L

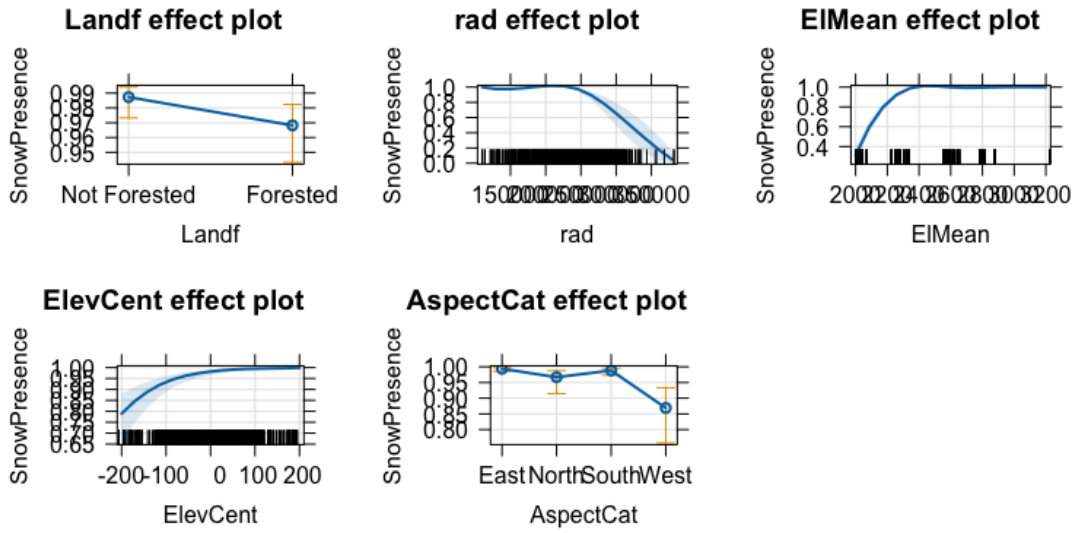
Run the following code to get re-started with the data set.

```
data(snowdepths)
snowdepths <- snowdepths %>%
  mutate(AspectCat = factor(case_when(
    aspect %in% (0:45)~ "North",
    aspect %in% (315:360)~ "North",
    aspect %in% 45:(90+45) ~ "East",
    aspect %in% (90+45):(180+45) ~ "South",
    aspect %in% (180+45):315 ~ "West"
  )),
  SnowPresence = factor(case_when(
    snow == 0 ~ "None",
    snow > 0 ~ "Some"
  )),
  Landf = factor(cover)
```

```
  )
levels(snowdepths$Landf) <- c("Not Forested", "Forested")

snowdepths <- snowdepths %>% mutate(ElMean = ave(elev, strata),
      ElevCent = elev - ElMean)

glm2 <- glm(SnowPresence ~ Landf + rad + ElMean + ElevCent + AspectCat,
            data = snowdepths, family = binomial)
summary(glm2)

##
## Call:
## glm(formula = SnowPresence ~ Landf + rad + ElMean + ElevCent +
##     AspectCat, family = binomial, data = snowdepths)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.501e+01  2.070e+00  -7.252 4.10e-13
## LandfForested  -9.386e-01  2.584e-01  -3.632 0.000281
## rad            -4.797e-05  5.572e-06  -8.610  < 2e-16
## ElMean          1.344e-02  1.232e-03  10.906  < 2e-16
## ElevCent        1.332e-02  1.844e-03   7.225 5.00e-13
## AspectCatNorth -1.694e+00  4.430e-01  -3.823 0.000132
## AspectCatSouth -6.353e-01  3.161e-01  -2.010 0.044460
## AspectCatWest  -3.175e+00  5.217e-01  -6.085 1.16e-09
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1072.4  on 1016  degrees of freedom
## Residual deviance:  493.3  on 1009  degrees of freedom
## AIC: 509.3
##
## Number of Fisher Scoring iterations: 8

plot(allEffects(glm2), grid = T, type="link")
```




```
plot(allEffects(glm2), grid = T, type = "response")
```



Q1 (repeat of Q11 from Lab 13) Interpret the `Landf` slope coefficient on the odds scale from `glm2` fit above.

```
1/exp(-9.386e-01)

## [1] 2.5564

confint(glm2)

##                          2.5 %        97.5 %
## (Intercept)    -1.924150e+01 -1.110907e+01
## LandfForested  -1.455363e+00 -4.395598e-01
## rad            -5.931315e-05 -3.743969e-05
## ElMean          1.117051e-02  1.600690e-02
## ElevCent        9.839887e-03  1.707647e-02
```

```
## AspectCatNorth -2.572461e+00 -8.305015e-01
## AspectCatSouth -1.261195e+00 -1.905017e-02
## AspectCatWest  -4.251595e+00 -2.195063e+00
```

```
1/exp(-1.455363e+00)
```

```
## [1] 4.286039
```

```
1/exp(-4.395598e-01)
```

```
## [1] 1.552024
```

The estimated odds of finding some snow in forested areas is 2.55 times higher (95% CI: 1.55, 4.28) than non-forested areas, after controlling for rad, ElMean, ElevCent, AspectCatNorth, AspectCatSouth, and AspectCatWest.

2) Check for multicollinearity in this `glm2` using the `vif` function. Report the results on using the rules of thumb and then specifically for the most impacted variable.

```
vif(glm2)
```

```
##               GVIF Df GVIF^(1/(2*Df))
## Landf     1.323570  1        1.150465
## rad       2.811005  1        1.676605
## ElMean    2.930991  1        1.712014
## ElevCent  1.458081  1        1.207510
## AspectCat 2.989578  3        1.200241
```

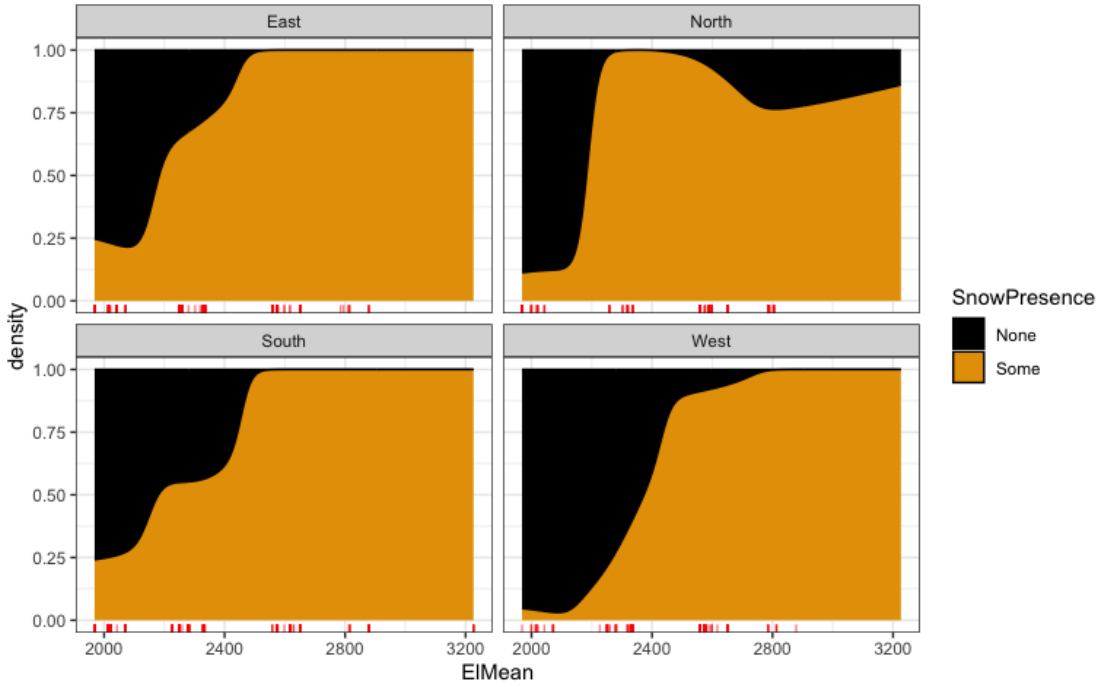
```
sqrt(vif(glm2))
```

```
##               GVIF       Df GVIF^(1/(2*Df))
## Landf     1.150465 1.000000        1.072597
## rad       1.676605 1.000000        1.294838
## ElMean    1.712014 1.000000        1.308439
## ElevCent  1.207510 1.000000        1.098868
## AspectCat 1.729040 1.732051        1.095555
```

All vifs are smaller than the rule of thumb VIF cutoff of 5. The most impacted variable ElMean, which has a VIF of 2.93. This suggests that the standard error is 1.712 times higher than would be expected without multicolinearity.

3) The following code makes a plot to visualize the response versus mean elevation by AspectCat. Modify the bandwidth to smooth but not oversmooth the density curves in the plots. Then interpret the relationship between mean elevation and snow presence and how it might change based on the aspect of the sites.

```
snowdepths %>% ggplot(aes(x = ElMean, fill = SnowPresence))  +
  geom_density(position='fill', bw = 70) +
  scale_fill_colorblind()  +
  geom_rug(aes(x = ElMean), alpha = 0.5, col = "red") +
  facet_wrap(vars(AspectCat))
```



Generally, odds of finding snow increases with increasing elevation. The north aspect has the lowest threshold for odds of finding snow increasing, and also shows a possible decrease in odds of finding snow at the highest elevations. There are very few observations in the high elevation north aspect category part of the dataset, so we have less information to support that possible trend. The west aspect has the highest threshold elevation for increased odds of finding snow.

4) We can do tests in GLMs using the z-statistics (distribution is the standard normal or just z with no DF to report) reported in the model summary or using `Anova` on the models to get tests similar to Type II ANOVA F-tests. The multi-degree of freedom tests use the Chi-squared distribution to find p-values but the hypotheses and interpretations otherwise match our previous work. Run `Anova` on the initial model and replace the censored parts of the following sentence to report the results from it for `AspectCat`.

```
Anova(glm2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: SnowPresence
##           LR Chisq Df Pr(>Chisq)
## Landf        13.84  1  0.0001994
## rad         104.02  1  < 2.2e-16
## ElMean      392.55  1  < 2.2e-16
## ElevCent     69.28  1  < 2.2e-16
## AspectCat    54.63  3  8.244e-12
```

- There is strong evidence against the null hypothesis of no difference in snow presence ($\chi^2_3 = 54.63$, p-value < 0.001) controlled for Landf, rad, ElMean, and

ElevCent so we would conclude that there is a difference snow presence rates across the different levels of AspectCat.

5) Replace the censored parts of the size sentence for the mean elevation predictor on the odds scale.

- For two sites that differ by 1 meter in the mean elevation of the site but are otherwise similar, the estimated mean odds of a site having snow is 1.013 times as much (95% profile likelihood CI from 1.011 to 1.016), controlled for Landf, rad, ElevCent, and AspectCat.

```
summary(glm2)
```

```
##
## Call:
## glm(formula = SnowPresence ~ Landf + rad + ElMean + ElevCent +
##     AspectCat, family = binomial, data = snowdepths)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.501e+01  2.070e+00  -7.252 4.10e-13
## LandfForested  -9.386e-01  2.584e-01  -3.632 0.000281
## rad            -4.797e-05  5.572e-06  -8.610  < 2e-16
## ElMean          1.344e-02  1.232e-03  10.906  < 2e-16
## ElevCent        1.332e-02  1.844e-03   7.225 5.00e-13
## AspectCatNorth -1.694e+00  4.430e-01  -3.823 0.000132
## AspectCatSouth -6.353e-01  3.161e-01  -2.010 0.044460
## AspectCatWest  -3.175e+00  5.217e-01  -6.085 1.16e-09
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1072.4  on 1016  degrees of freedom
## Residual deviance:  493.3  on 1009  degrees of freedom
## AIC: 509.3
##
## Number of Fisher Scoring iterations: 8
```

```
exp(1.344e-02)
```

```
## [1] 1.013531
```

```
confint(glm2)
```

```
##                        2.5 %        97.5 %
## (Intercept)    -1.924150e+01 -1.110907e+01
## LandfForested  -1.455363e+00 -4.395598e-01
## rad            -5.931315e-05 -3.743969e-05
## ElMean          1.117051e-02  1.600690e-02
## ElevCent        9.839887e-03  1.707647e-02
## AspectCatNorth -2.572461e+00 -8.305015e-01
## AspectCatSouth -1.261195e+00 -1.905017e-02
## AspectCatWest  -4.251595e+00 -2.195063e+00
```

```
exp(1.117051e-02)

## [1] 1.011233

exp(1.600690e-02)

## [1] 1.016136
```

6) In GLMs, we can also do pairwise comparisons using a version of Tukey-Kramer comparisons using emmeans. The results are on the link (here logit or log-odds) scale but can be translated to be on the odds scale. The following code generates pairwise comparisons among the levels of Aspect and provides an interpretation of the size for the East versus West facing sites on the odds scale. Modify the provided sentence to report the size for the North to West comparison.

- The estimated mean odds of encountering snow in an North facing site are 4.397 times as much as for a West facing site (95% familywise-CI from 0.927 to 20.843), controlled for land cover, radiation, elevation strata mean and variation of the site around the strata mean.

```
res1 <- emmeans(glm2, pairwise ~ AspectCat, adjust = "tukey")
confint(res1)

## $emmeans
##  AspectCat emmean    SE  df asymp.LCL asymp.UCL
##  East        4.97 0.416 Inf      4.15      5.78
##  North       3.27 0.504 Inf      2.29      4.26
##  South       4.33 0.422 Inf      3.51      5.16
##  West        1.79 0.380 Inf      1.05      2.54
##
## Results are averaged over the levels of: Landf
## Results are given on the logit (not the response) scale.
## Confidence level used: 0.95
##
## $contrasts
##  contrast      estimate    SE  df asymp.LCL asymp.UCL
##  East - North     1.694 0.443 Inf     0.556     2.832
##  East - South     0.635 0.316 Inf    -0.177     1.447
##  East - West      3.175 0.522 Inf     1.834     4.515
##  North - South   -1.058 0.500 Inf    -2.343     0.226
##  North - West     1.481 0.606 Inf    -0.075     3.037
##  South - West     2.539 0.523 Inf     1.197     3.882
##
## Results are averaged over the levels of: Landf
## Results are given on the log odds ratio (not the response) scale.
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 4 estimates

summary(res1)

## $emmeans
##  AspectCat emmean    SE  df asymp.LCL asymp.UCL
```

```
##  East        4.97 0.416 Inf      4.15      5.78
##  North       3.27 0.504 Inf      2.29      4.26
##  South       4.33 0.422 Inf      3.51      5.16
##  West        1.79 0.380 Inf      1.05      2.54
##
## Results are averaged over the levels of: Landf
## Results are given on the logit (not the response) scale.
## Confidence level used: 0.95
##
## $contrasts
##  contrast      estimate    SE  df z.ratio p.value
##  East - North     1.694 0.443 Inf   3.823  0.0008
##  East - South     0.635 0.316 Inf   2.010  0.1843
##  East - West      3.175 0.522 Inf   6.085  <.0001
##  North - South   -1.058 0.500 Inf  -2.117  0.1477
##  North - West     1.481 0.606 Inf   2.445  0.0689
##  South - West     2.539 0.523 Inf   4.859  <.0001
##
## Results are averaged over the levels of: Landf
## Results are given on the log odds ratio (not the response) scale.
## P value adjustment: tukey method for comparing a family of 4 estimates
```

```
exp(3.175)
```

```
## [1] 23.92682
```

```
exp(c(1.834, 4.515))
```

```
## [1]  6.258872 91.377566
```

```
exp(1.481)
```

```
## [1] 4.397341
```

```
exp(c(-0.075 , 3.037))
```

```
## [1]  0.9277435 20.8426215
```

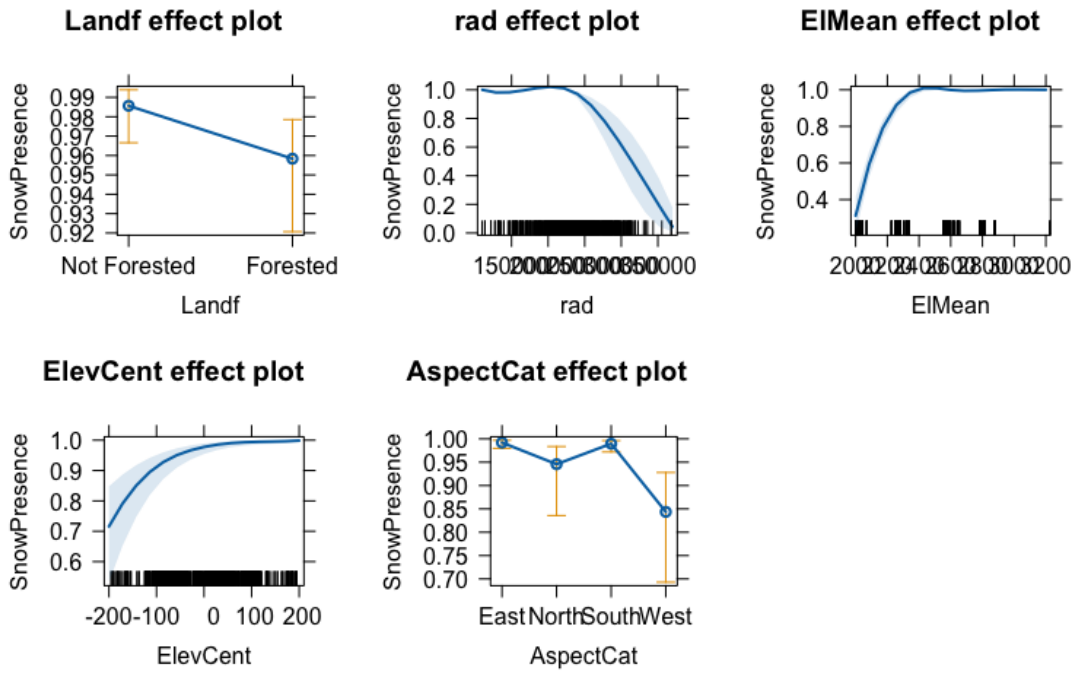
Predicting Snow Presence

If we were interested in assessing the prediction error in this situation, we would want to split the data set into training and test data sets. For time, we won't do any model selection within the training data and just assess the predictive performance of our initial model. The following code splits the data set and re-fits the model just to the training data and does some work to visualize that new version of the model.

7) What are the units of the y-axis in the `enhanced_stripcharts` that the code below produces?

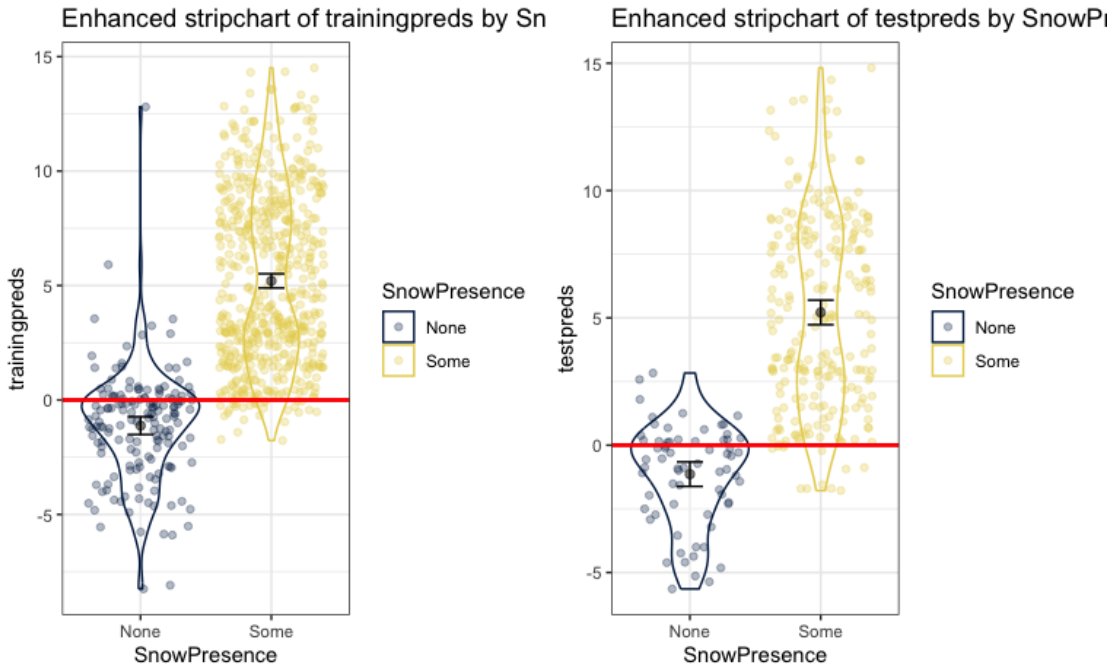
```
set.seed(123)
trainingD <- snowdepths %>% slice_sample(prop = 0.7)
```

```
testD <- anti_join(x = snowdepths, y = trainingD)
dim(trainingD)

## [1] 711  15

dim(testD)

## [1] 306  15

glm2_train <- glm(SnowPresence ~ Landf + rad + ElMean + ElevCent + AspectCat,
            data = trainingD, family = binomial)

glm2$coefficients

##    (Intercept)  LandfForested            rad         ElMean       ElevCent
##  -1.501208e+01  -9.385713e-01  -4.797149e-05   1.343573e-02   1.332197e-02
## AspectCatNorth AspectCatSouth  AspectCatWest
##  -1.693634e+00  -6.353198e-01  -3.174642e+00

glm2_train$coefficients

##    (Intercept)  LandfForested            rad         ElMean       ElevCent
##  -1.345715e+01  -1.094704e+00  -5.072821e-05   1.297768e-02   1.424236e-02
## AspectCatNorth AspectCatSouth  AspectCatWest
##  -1.973341e+00  -2.716770e-01  -3.145542e+00

plot(allEffects(glm2_train), type = "response")
```




```
trainingD <- trainingD %>% mutate(trainingpreds = predict(glm2_train, newdata
= trainingD))
```

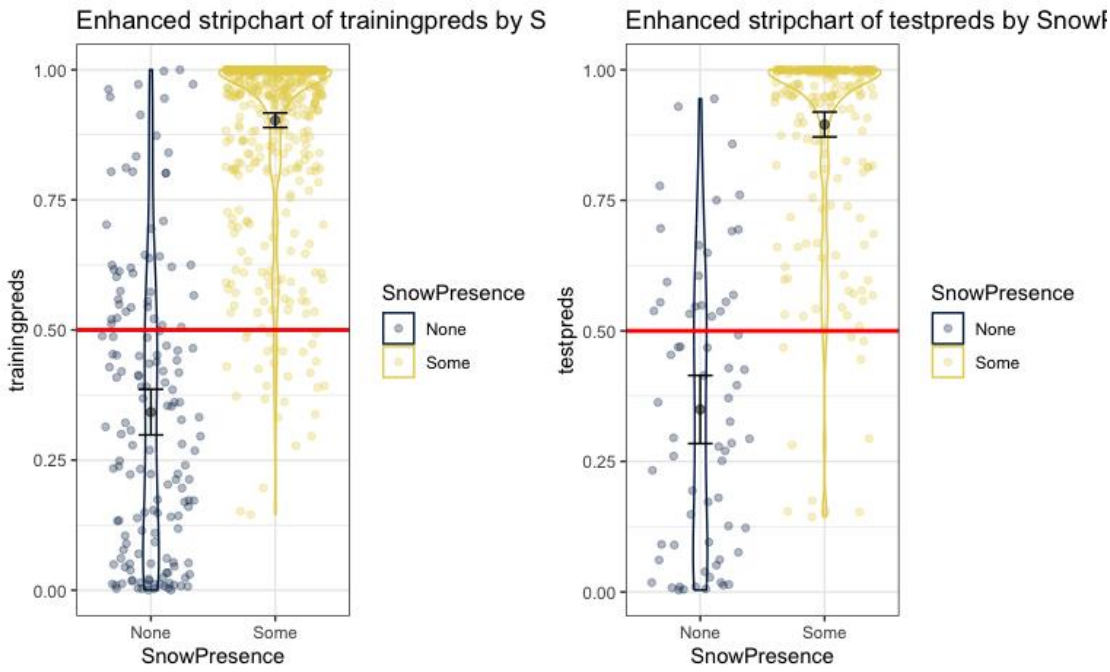
```
testD <- testD %>% mutate(testpreds = predict(glm2_train, newdata = testD))

p1 <- enhanced_stripchart(trainingpreds ~ SnowPresence, data = trainingD) +
geom_hline(yintercept = 0, col = "red", lwd = 1)
p2 <- enhanced_stripchart(testpreds ~ SnowPresence, data = testD) +
geom_hline(yintercept = 0, col = "red", lwd = 1)

p1 + p2
```



The units of the y-axis are log odds.

8) Modify the following code to make the `enhanced_stripcharts` for the predictions to be on the response scale instead of the link scale and change the `geom_hline` to be useful for considering prediction results on the response scale. No discussion.

```
#Modify where appropriate:
trainingD <- trainingD %>% mutate(trainingpreds = predict(glm2_train, newdata
= trainingD, type= "response"))

testD <- testD %>% mutate(testpreds = predict(glm2_train, newdata = testD,
type= "response"))

p1 <- enhanced_stripchart(trainingpreds ~ SnowPresence, data = trainingD) +
geom_hline(yintercept = 0.5, col = "red", lwd = 1)
p2 <- enhanced_stripchart(testpreds ~ SnowPresence, data = testD) +
geom_hline(yintercept = 0.5, col = "red", lwd = 1)

p1 + p2
```



9) Using a probability of 0.5 from the estimated model, what proportion of the test data are predicted to be in the "success" category?

```
library(caret)

p3 <- tally(testpreds>0.5 ~ 1, data = testD)
p3

##                 1
## testpreds > 0.5   1
##           TRUE  250
##           FALSE  56

250/(250+56)

## [1] 0.8169935
```

81.6% of the test data were predicted to be in the success category.

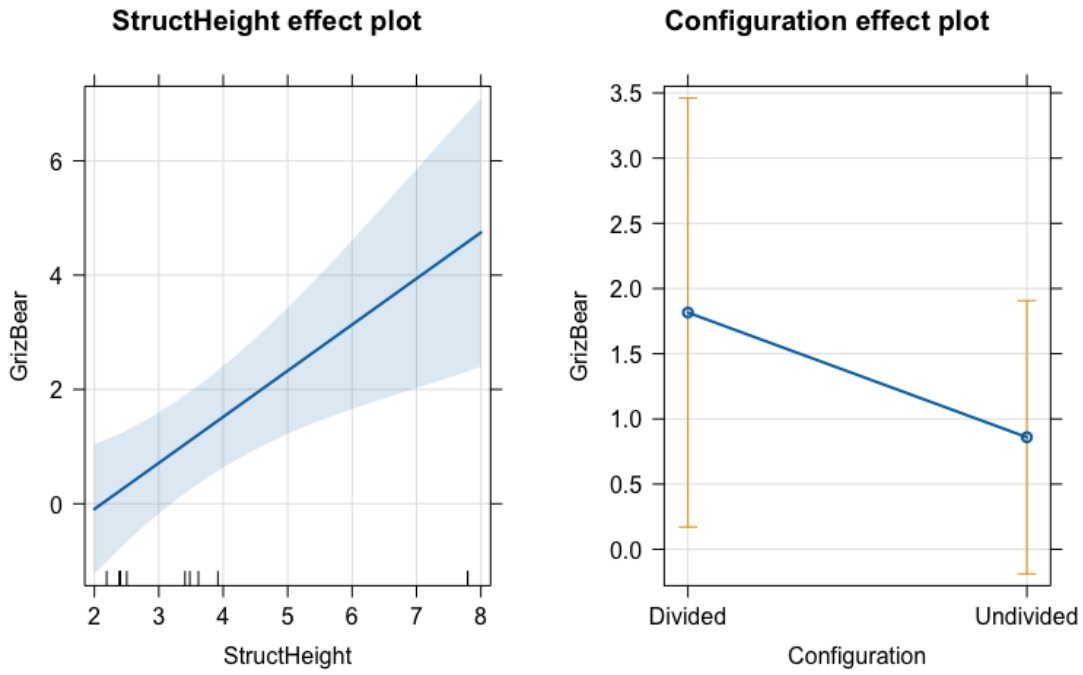
Part III: Grizzly Bear Counts:

In homework 2, we explored counts of grizzly bears using road crossing structures and you fit the following model:

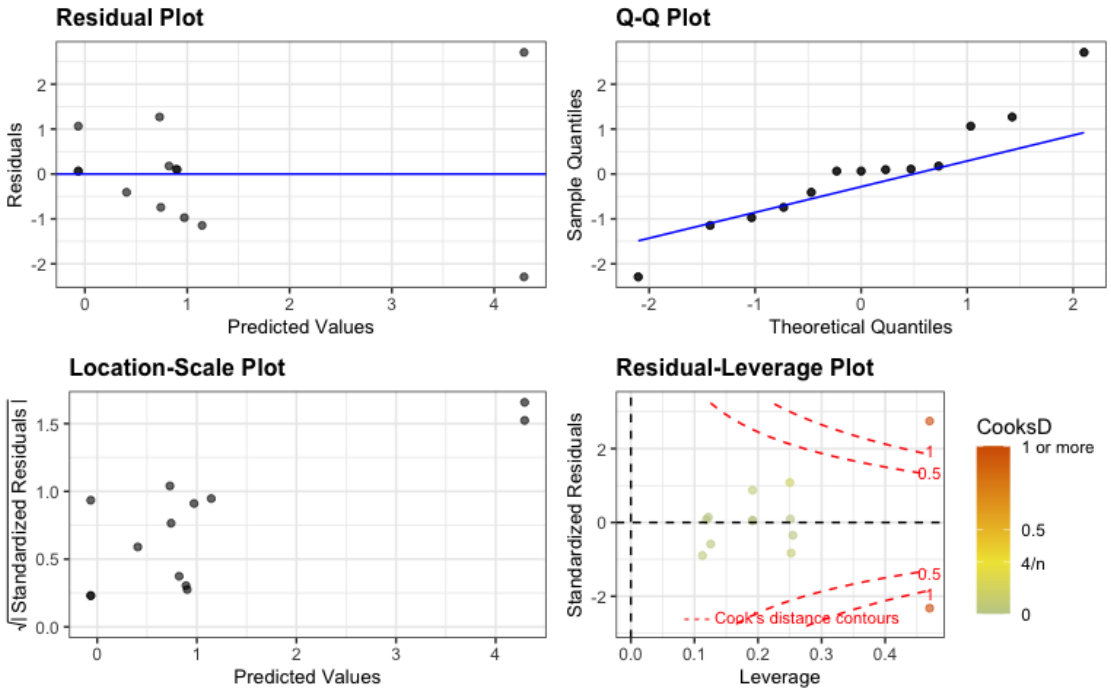
```
clev <- read_excel("clev_update.xlsx")
clev <- clev %>% mutate(Configuration = factor(Configuration))

lm1 <- lm(GrizBear ~ StructHeight + Configuration, data = clev)
summary(lm1)
```

```
##
## Call:
## lm(formula = GrizBear ~ StructHeight + Configuration, data = clev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.29123 -0.74175  0.06495  0.17758  2.70877
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1.0449     0.8378  -1.247  0.24073
## StructHeight               0.8067     0.2214   3.643  0.00451
## ConfigurationUndivided    -0.9561     0.9175  -1.042  0.32191
##
## Residual standard error: 1.355 on 10 degrees of freedom
## Multiple R-squared:  0.5796, Adjusted R-squared:  0.4955
## F-statistic: 6.894 on 2 and 10 DF,  p-value: 0.01313
```

```
plot(allEffects(lm1), grid = T)
```




```
resid_panel(lm1, "R")
```




```
vif(lm1)

##  StructHeight Configuration
##      1.269161      1.269161
```

- $GrizzlyCount_i \sim N(\mu_i, \sigma^2)$
- $\hat{\mu}_i = -1.045 + 0.8067 StructHeight_i - 0.956 I_{Config=Undivided,i}$
- where i = 1, …, 13 structures and $I_{Config=Undivided,i}$ is 1 for undivided structures and 0 otherwise.

10) Here is a similar analysis using a Poisson GLM that more appropriately models the counts of grizzly bears that used the structures in the time period of the study. Write out the estimated model. Hint: see the introductory and Ch. 22 lecture notes.

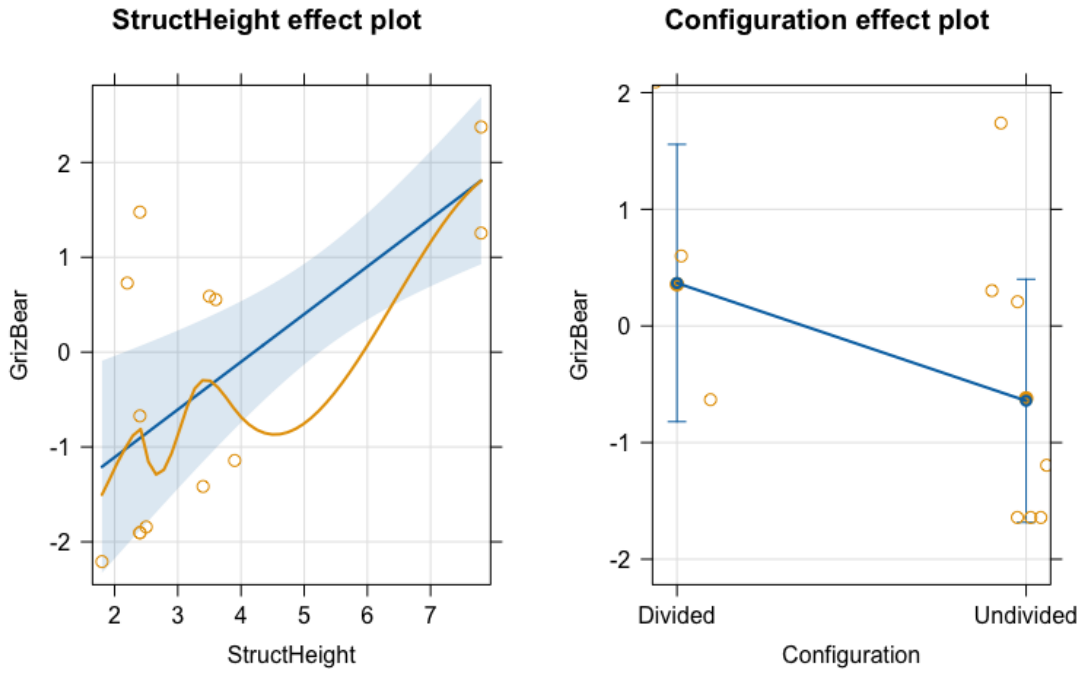
```
glm1 <- glm(GrizBear ~ StructHeight + Configuration, data = clev, family =
"poisson")

summary(glm1)

##
## Call:
## glm(formula = GrizBear ~ StructHeight + Configuration, family = "poisson",
##     data = clev)
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.4154     0.6629  -2.135 0.032754
```

```
## StructHeight              0.5029     0.1442   3.488 0.000487
## ConfigurationUndivided  -1.0097     0.9035  -1.118 0.263741
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 28.495  on 12  degrees of freedom
## Residual deviance: 12.484  on 10  degrees of freedom
## AIC: 35.519
##
## Number of Fisher Scoring iterations: 5
```

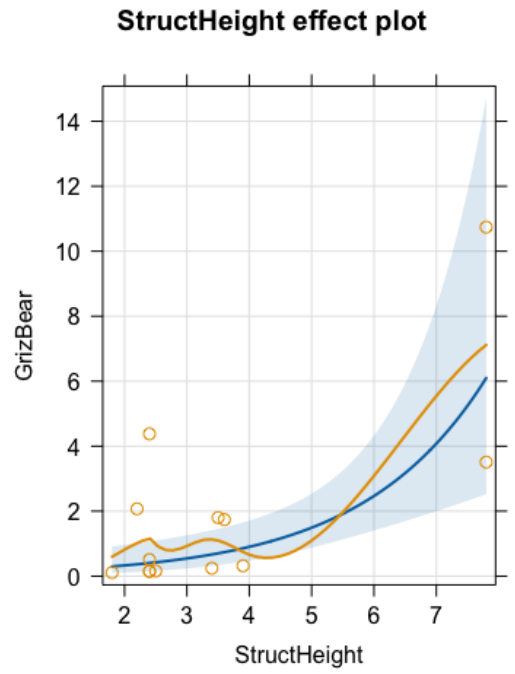
```
confint(glm1)
```

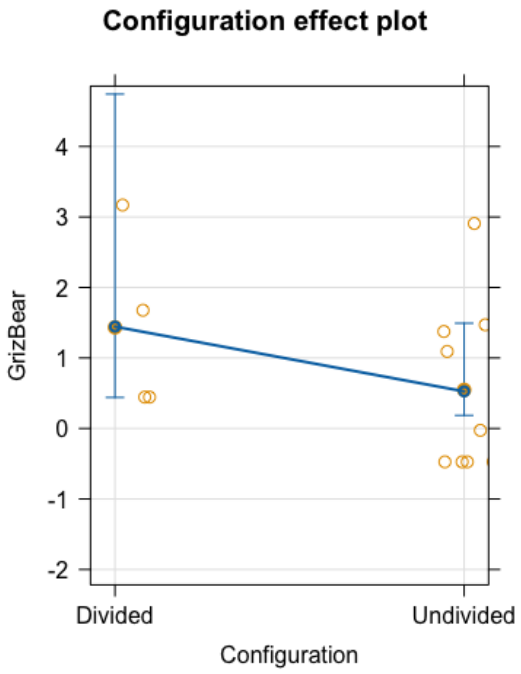
```
##                               2.5 %     97.5 %
## (Intercept)            -2.9364569 -0.2660486
## StructHeight            0.2426123  0.8291659
## ConfigurationUndivided -2.9365890  0.7914657
```

```
plot(allEffects(glm1, residuals = T), grid = T, type = "link")
```

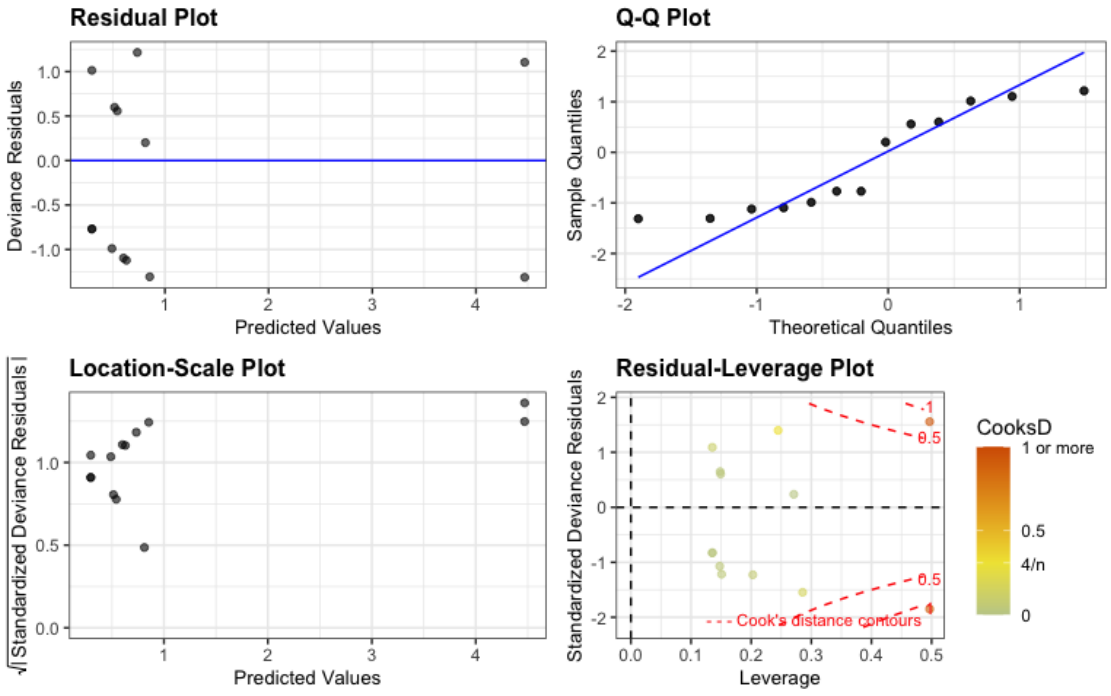



```
plot(allEffects(glm1, residuals = T), grid = T, type = "response")
```






```
resid_panel(glm1, "R")
```




```
vif(glm1)

##  StructHeight Configuration
##      1.959015      1.959015
```

- $GrizzlyCount_i \sim Poisson(\mu_i, \sigma^2)$

- $log(*i[GrizzlyBears|StructureHeight,Configuration]) = -1.415 + 0.5029(StructureHeight) - 1.0097I*{Configuration=Undivided} $
- where i = 1, …, 13 structures and $I_{Configuration=Undivided,i}$ is 1 for undivided structures and 0 otherwise.