Lab 6

DO NOT INCLUDE NAMES - Just add names in gradescope

# Rules

In groups of 2 or 3, complete the following.

# Modeling Snow Density

We will be using the data set from Wetlaufer, Hendrikx, and Marshall (2016) - study where they explored the relationship between snow density () or snow depth (snow, mm) or snow water equivalent (SWE, mm) with a suite of predictor variables. In this lab we will condition on there being snow present and then try to model snow density with models that will compete with what they discuss in Tables 3 and 4. We will be interested in using elev (Elevation, m), Land (forest cover with 0 = unforested and 10 = forested), rad (Potential Solar radiation, ), 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 degrees)), and angle (angle of slope in degrees with 0 being flat) as 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, as we will revisit this.

* 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 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),  
 Landf = fct\_recode(Landf,  
 "Not Forested" = "0",  
 "Forested" = "10")  
 )  
  
favstats(aspect ~ 1, data = snowdepths)

## 1 min Q1 median Q3 max mean sd n missing  
## 1 1 0 84 153 228 359 159.061 93.78348 1017 0

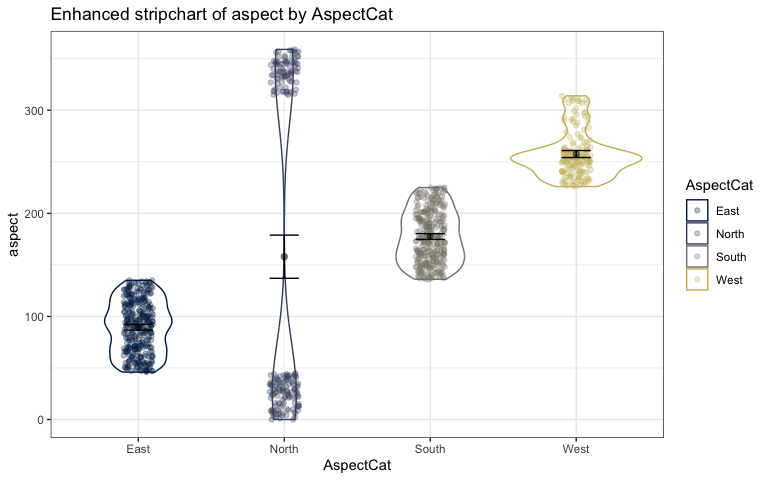
favstats(aspect ~ AspectCat, data = snowdepths)

## AspectCat min Q1 median Q3 max mean sd n missing  
## 1 East 46 68.00 91 111.25 135 89.51562 25.38308 320 0  
## 2 North 0 24.25 41 335.75 359 157.92056 155.45378 214 0  
## 3 South 136 157.00 176 198.00 225 177.47604 25.11551 313 0  
## 4 West 226 241.00 253 263.75 314 257.50000 22.41836 170 0

**1) Before we dig into modeling, we need to consider how we are going to handle the aspect variable. Consider the sample mean of 159 and other summary statistics for the entire data set or the similar results broken down by AspectCat. Make an enhanced stripchart to help you to explain the favstats output for the AspectCat variable that I created from aspect. Why are the conventional summary statistics misleading with a variable such as aspect?**

Aspect here is indicating the face direction of the surface, and it coded as certain values out of 360 degrees. Think of it like looking around a circle as different degree angles. The AspectCat variable splits these values into ranges that represent the 4 cardinal directions. Conventional summary statistics are misleading with variables such as aspect because they are treating it as a continuous variable instead of a circular or categorical one.

enhanced\_stripchart(data = snowdepths, aspect ~ AspectCat)



**2) angle is also measured in degrees for the steepness of the slope. Create summary statistics for angle and then discuss how things are different for working with this variable as compared to aspect?**

Angle is not circular in nature like aspect is, and could theoretically be any range of values from flat to 90 degrees. We are looking at a range of values for this variable instead of identifiers like the values for aspect that signal direction.

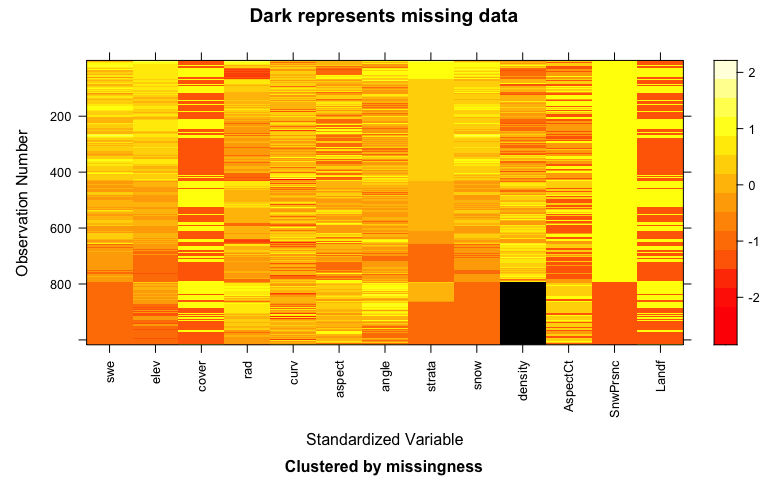
favstats(angle ~ 1, data = snowdepths)

## 1 min Q1 median Q3 max mean sd n missing  
## 1 1 1 8 12 19 51 14.39823 8.792653 1017 0

**3) Explain the missing data pattern and discuss the sample size before and after cleaning based on the code provided.**

It looks like there were a good amount of missing density observations that were then cleaned from the dataset. Before any cleaning there were 1017 observations of the 13 variables, and after cleaning this number dropped to 793 observations because of the removal of observations missing density.

library(mi)  
snowdepths %>% as.data.frame() %>% missing\_data.frame() %>% image()



tdf <- missing\_data.frame(data.frame(snowdepths))  
table(tdf@patterns)

##   
## nothing density   
## 793 224

snowdepthsR <- snowdepths %>% drop\_na(density)  
  
dim(snowdepths)

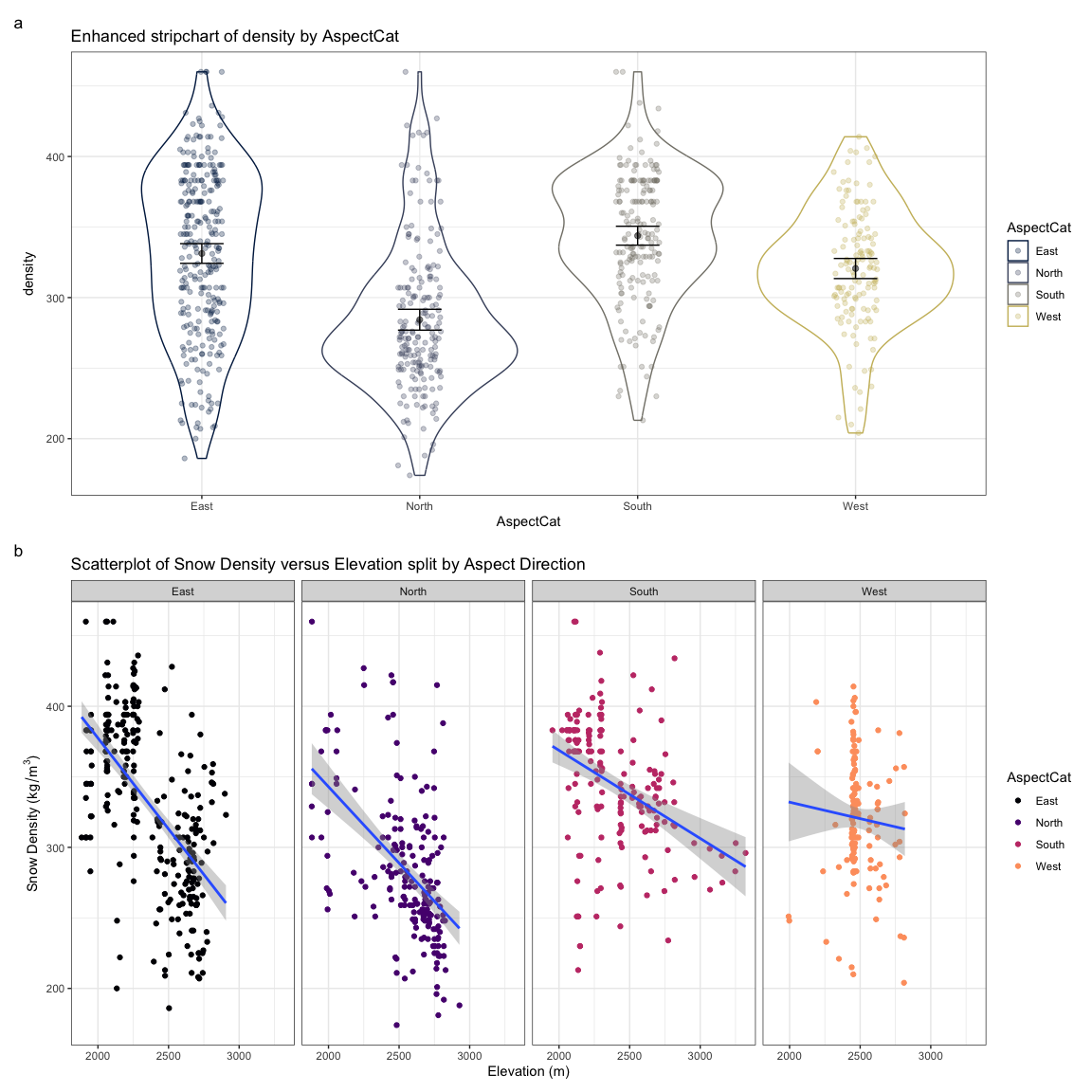
## [1] 1017 13

dim(snowdepthsR)

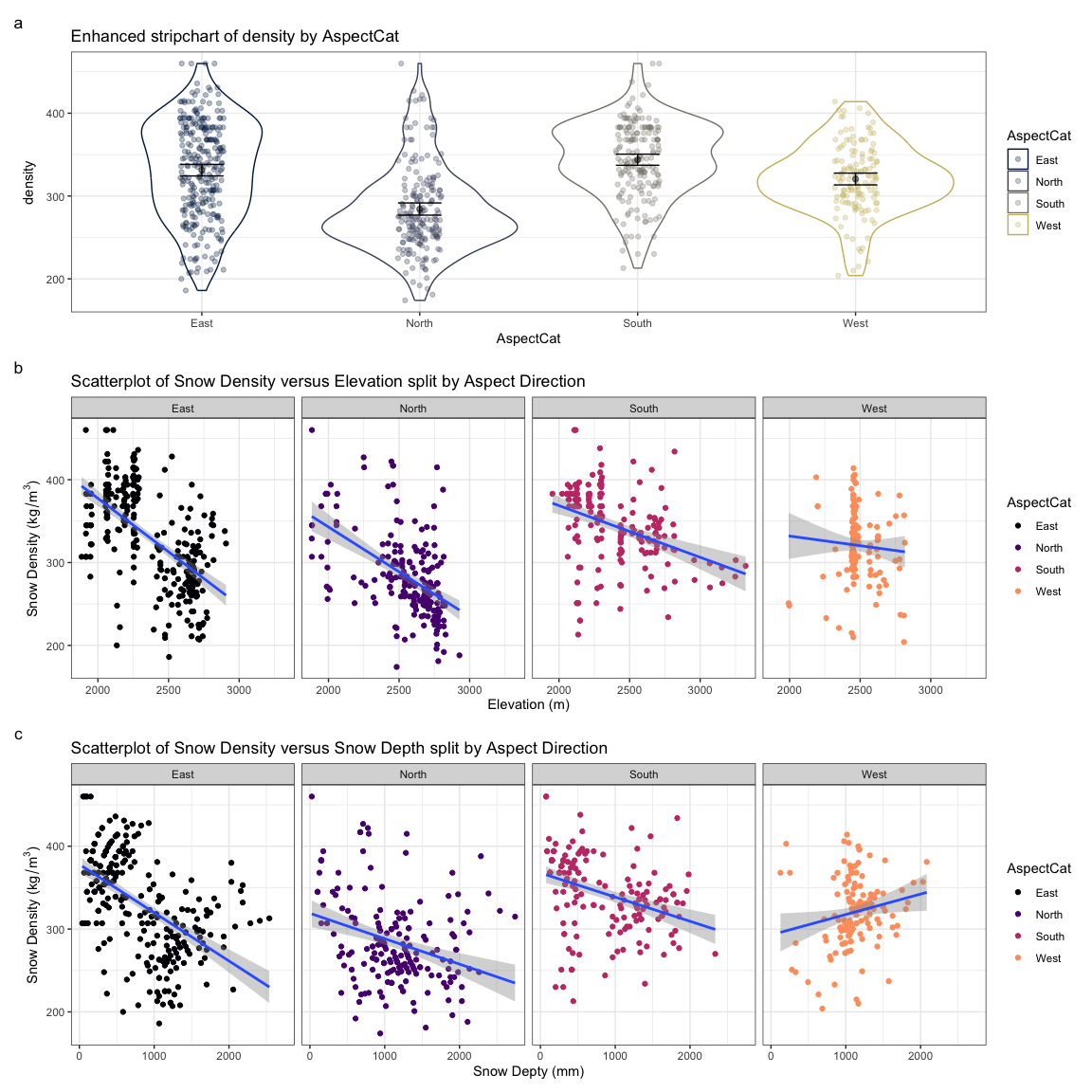
## [1] 793 13

**4) The code below creates two plots for the snow density response variable in the cleaned data set. Fix the title and labels to include units for the variables in p2. Then add a third plot (called p3) to the set of provided plots to explore snow density (density) versus snow depth (snow) by AspectCat, also with improved axis labels and a title. No discussion.**

p1 <- enhanced\_stripchart(density ~ AspectCat, data = snowdepthsR)  
p2 <- snowdepthsR %>% ggplot(aes(x = elev, y= density)) +  
 geom\_point(aes(color = AspectCat)) +  
 scale\_color\_viridis\_d(end = 0.8, option = "A") +  
 geom\_smooth(method = "lm") +  
 facet\_wrap(~AspectCat, nrow = 1) +  
 labs(title = "Scatterplot of Snow Density versus Elevation split by Aspect Direction",  
 x = "Elevation (m)",  
 y = bquote("Snow Density (" \* kg/m^3 \* ")"))  
  
p1 / p2 + plot\_annotation(tag\_levels = "a")



p3 <- snowdepthsR %>% ggplot(aes(x = snow, y= density)) +  
 geom\_point(aes(color = AspectCat)) +  
 scale\_color\_viridis\_d(end = 0.8, option = "A") +  
 geom\_smooth(method = "lm") +  
 facet\_wrap(~AspectCat, nrow = 1) +  
 labs(title = "Scatterplot of Snow Density versus Snow Depth split by Aspect Direction",  
 x = "Snow Depty (mm)",  
 y = bquote("Snow Density (" \* kg/m^3 \* ")"))  
  
  
  
p1 / p2 / p3 + plot\_annotation(tag\_levels = "a")



lmA <- lm(density ~ rad + snow + elev + AspectCat, data = snowdepthsR)

**5) Run Anova from the car package on lmA. Use the results to report an evidence sentence for AspectCat.**

**There is weak evidence against the null hypothesis of no difference in the true mean snow density across the aspect categories (F(3, 786) = 1.44, p-value = 0.23) controlled for radiation, snow depth, and elevation, so we would conclude that there isn't a difference after adjusting for those other variables and drop it from the model.**

Anova(lmA)

## Anova Table (Type II tests)  
##   
## Response: density  
## Sum Sq Df F value Pr(>F)  
## rad 126488 1 65.9686 1.77e-15  
## snow 14275 1 7.4451 0.006503  
## elev 261730 1 136.5022 < 2.2e-16  
## AspectCat 8289 3 1.4411 0.229477  
## Residuals 1507078 786

**6) We should assess model assumptions before we report any inferences from a model. Discuss the potential for a violation of the independence assumption in the previous model for snow density based on the design of the study and data collection. Hint: review the paper (Sections 2 and 3) for how they decided on their sampling locations.**

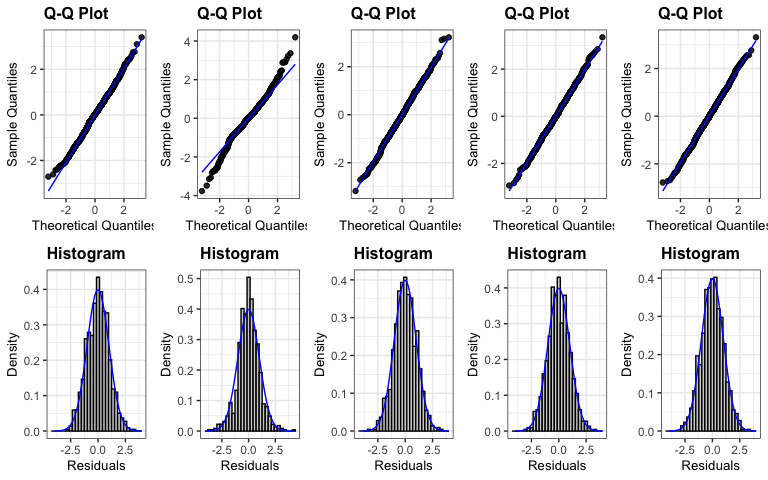
Based on their sampling and study design there are possible violations against the independence assumption. Sampling was conducted in a few distinct areas of West Fork Basin that were selected for their characteristics of different terrain types and defined as physiographic strata. Within these areas, sampling points were randomly selected at distances ranging from 30 to 400m. At each of these sampling points, three samples were taken 10m apart in triangular fashion. This cluster sampling may introduce violations to the assumption of independence as these points may show similar values and not be independent of each other.

**7) Run the following code on your model to generate a calibration set of plots for the residuals for assessing the normality of residuals assumption. (a) Which column contained the real residuals? (b) How easy was it to identify the real residuals versus versions of the residuals simulated from normal distributions (all the others)? (c) Use how clear it was to identify the real residuals from the simulated ones to discuss the strength of evidence against the normality of residuals assumption. Make sure you discuss the shape of the residual distribution if you think it doesn’t follow a normal distribution.**

The real residuals are seen in column 2. It was somewhat easy to identify the real residuals since they display a light-tailed distribution that is not present in the other 4 simulations. Based on this difference we would say that there is strong evidence against the assumption of normality.

set.seed(123)  
resid\_calibrate(lmA, nsim = 4, c("qq","hist"), identify = T, shuffle = T, coordfix = F)

## [1] "Real residuals are in column 2"

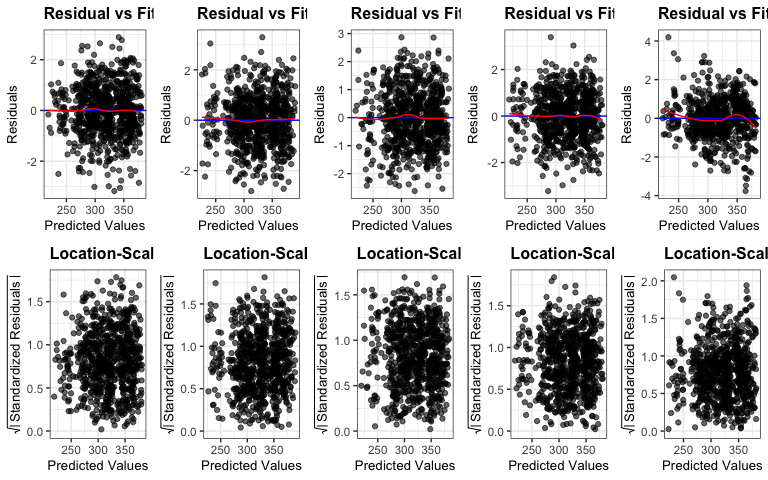


**8) There is no clear issue with linearity in the residuals vs fitted (more later on that with partial residuals). So we are safe to consider the residuals vs fitted and location-scale plots for assessing the equal variance of residuals assumption. Use the following resid\_calibrate code on your model to explore the equal variance assumption. Write a sentence or two to discuss the strength of evidence against the assumption of equal variance based on these results. When it is safe to consider, you should always discuss the Location-Scale plot along with the Residuals vs Fitted when assessing HOV.**

I would say that there is little evidence against the assumption of equal variance as the residuals vs fitted plot is difficult to identify from the simulated normal distributions.

set.seed(12345)  
resid\_calibrate(lmA, c("resid", "ls"), nsim = 4, identify = T, shuffle = T)

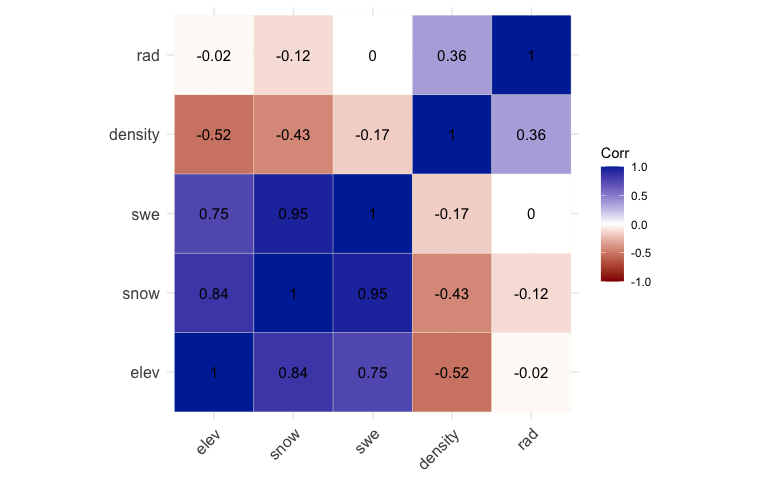
## [1] "Real residuals are in column 5"



**9) Two more predictors are added to the model (cover (whether location had trees or not) and swe (snow water equivalent)). Check for issues with multi-collinearity in lmA using the vif function (from car) and the check\_collinearity function (from performance). Which two variables are most impacted by multi-collinearity in lmB? How can you explain that?**

Snow Water Equivalent (SWE) and Snow Depth variables are most impacted by multi-collinearity. SWE measures the amount of water contained in the snowpack by representing the depth of water that would result if the snow were melted. It makes sense that this variable is highly correlated with snow depth as depth is used to calculate swe.

ggcorrplot(snowdepthsR %>% dplyr::select(density, rad, snow, elev, swe) %>% cor(), lab = T)



lmB <- lm(density ~ rad + snow + elev + AspectCat + cover + swe, data = snowdepthsR)  
vif(lmB)

## GVIF Df GVIF^(1/(2\*Df))  
## rad 2.164159 1 1.471108  
## snow 20.715254 1 4.551401  
## elev 4.136097 1 2.033740  
## AspectCat 2.350019 3 1.153042  
## cover 1.091445 1 1.044722  
## swe 13.970419 1 3.737702

library(performance)  
check\_collinearity(lmB)

## # Check for Multicollinearity  
##   
## Low Correlation  
##   
## Term VIF VIF 95% CI Increased SE Tolerance Tolerance 95% CI  
## rad 2.16 [ 1.96, 2.42] 1.47 0.46 [0.41, 0.51]  
## elev 4.14 [ 3.67, 4.68] 2.03 0.24 [0.21, 0.27]  
## AspectCat 2.35 [ 2.12, 2.63] 1.53 0.43 [0.38, 0.47]  
## cover 1.09 [ 1.04, 1.22] 1.04 0.92 [0.82, 0.96]  
##   
## High Correlation  
##   
## Term VIF VIF 95% CI Increased SE Tolerance Tolerance 95% CI  
## snow 20.72 [18.13, 23.69] 4.55 0.05 [0.04, 0.06]  
## swe 13.97 [12.25, 15.96] 3.74 0.07 [0.06, 0.08]

**10) How can you explain the following change in slope coefficients for snow and swe across the following three models?**

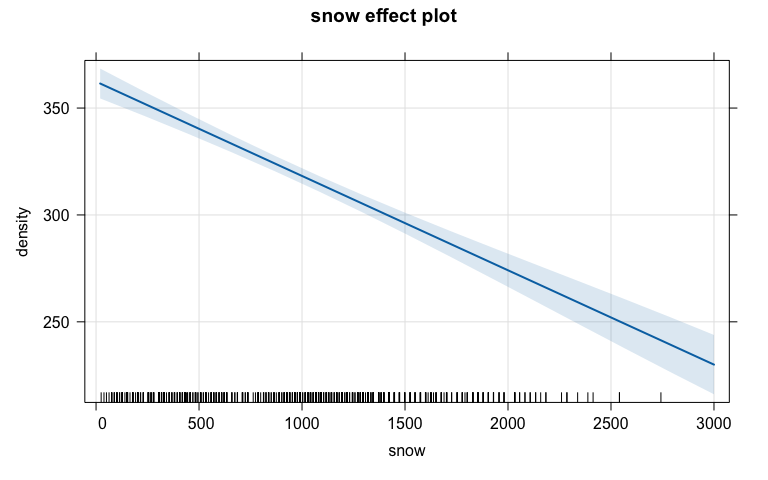
In the models with only snow depth or swe as the only predictor variables, slope coefficients were -0.044(depth) and -0.052(swe). These negative slope values suggest that as snow depth or swe increased, density decreased based on one predictor.

In the additive model, slope coefficient for depth was -0.285614 and 0.767854 for swe. The negative coefficient for snow is now larger because it reflects the independent effect of snow, after removing the influence of SWE. SWE now seems to have a positive effect on density in the presence of snow. This is due to multicollinearity and their shared variance.

lmDepth <- lm(density ~ snow, data = snowdepthsR)  
summary(lmDepth)

##   
## Call:  
## lm(formula = density ~ snow, data = snowdepthsR)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -146.886 -35.635 5.739 34.281 152.344   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 362.365906 3.618679 100.14 <2e-16  
## snow -0.044128 0.003322 -13.29 <2e-16  
##   
## Residual standard error: 51.01 on 791 degrees of freedom  
## Multiple R-squared: 0.1824, Adjusted R-squared: 0.1814   
## F-statistic: 176.5 on 1 and 791 DF, p-value: < 2.2e-16

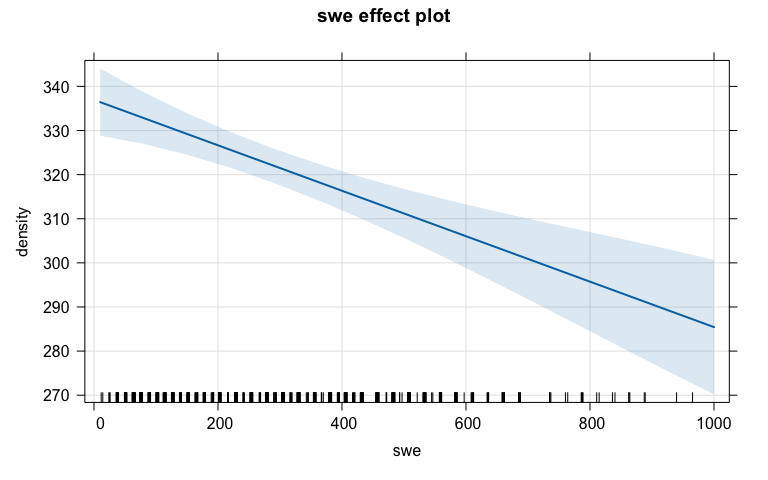
plot(allEffects(lmDepth), grid = T)



lmSWE <- lm(density ~ swe, data = snowdepthsR)  
summary(lmSWE)

##   
## Call:  
## lm(formula = density ~ swe, data = snowdepthsR)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -153.782 -40.697 4.303 43.760 141.556   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 336.95121 3.96880 84.900 < 2e-16  
## swe -0.05151 0.01094 -4.708 2.96e-06  
##   
## Residual standard error: 55.64 on 791 degrees of freedom  
## Multiple R-squared: 0.02725, Adjusted R-squared: 0.02602   
## F-statistic: 22.16 on 1 and 791 DF, p-value: 2.958e-06

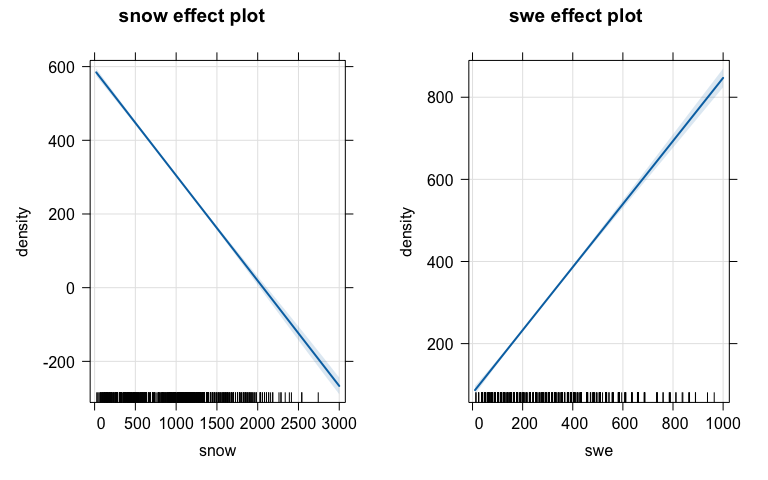
plot(allEffects(lmSWE), grid = T)



lmC <- lm(density ~ snow + swe, data = snowdepthsR)  
summary(lmC)

##   
## Call:  
## lm(formula = density ~ snow + swe, data = snowdepthsR)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -99.772 -11.555 -3.861 14.672 109.771   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 348.591457 1.925461 181.04 <2e-16  
## snow -0.285614 0.005582 -51.17 <2e-16  
## swe 0.767854 0.016858 45.55 <2e-16  
##   
## Residual standard error: 26.81 on 790 degrees of freedom  
## Multiple R-squared: 0.7745, Adjusted R-squared: 0.774   
## F-statistic: 1357 on 2 and 790 DF, p-value: < 2.2e-16

plot(allEffects(lmC), grid = T)



**11) Report any resources outside your group and the course provided materials that you used and how they impacted your answers.**

NONE