Assignment 7: Correlation and Regression

Asa Hayes

19 April, 2021

Instructions

Download the assignment07.Rmd file from Canvas and open it in RStudio. Complete this assignment by filling in the answers below in the R Markdown Notebook document.

Data

In this assignment, you will apply correlation and regression analysis to environmental variables located at the PUF Lands CRIS Survey drilling pad measurements that were collected for Assignment #4. Similar to Assignment #6, pads that are within 500 meters of any other pads have been removed and geographic latitude/longitude coordinates have been converted to UTM Zone 14 coordinates.

The drilling pad data are contained within a CSV file called "CRIS_AI_elev.csv", located at the following URL:
http://people.tamu.edu/~geoallen/courses/312/CRIS_AI_elev.csv
This CSV contains the following variables: (1) approach; (2) x; (3) y; (4) area_m2 (the area of the pad); (5) aridity (mean aridity index at the pad from CGIAR); (6) elev_m (pad elevation in meters above sea level from the 15 Arcsecond HydroSHEDS DEM).

Deliverables

Please submit to Canvas the following items:
1. An HTML (or Word or PDF) file knitted from the .Rmd file
2. A completed .Rmd file

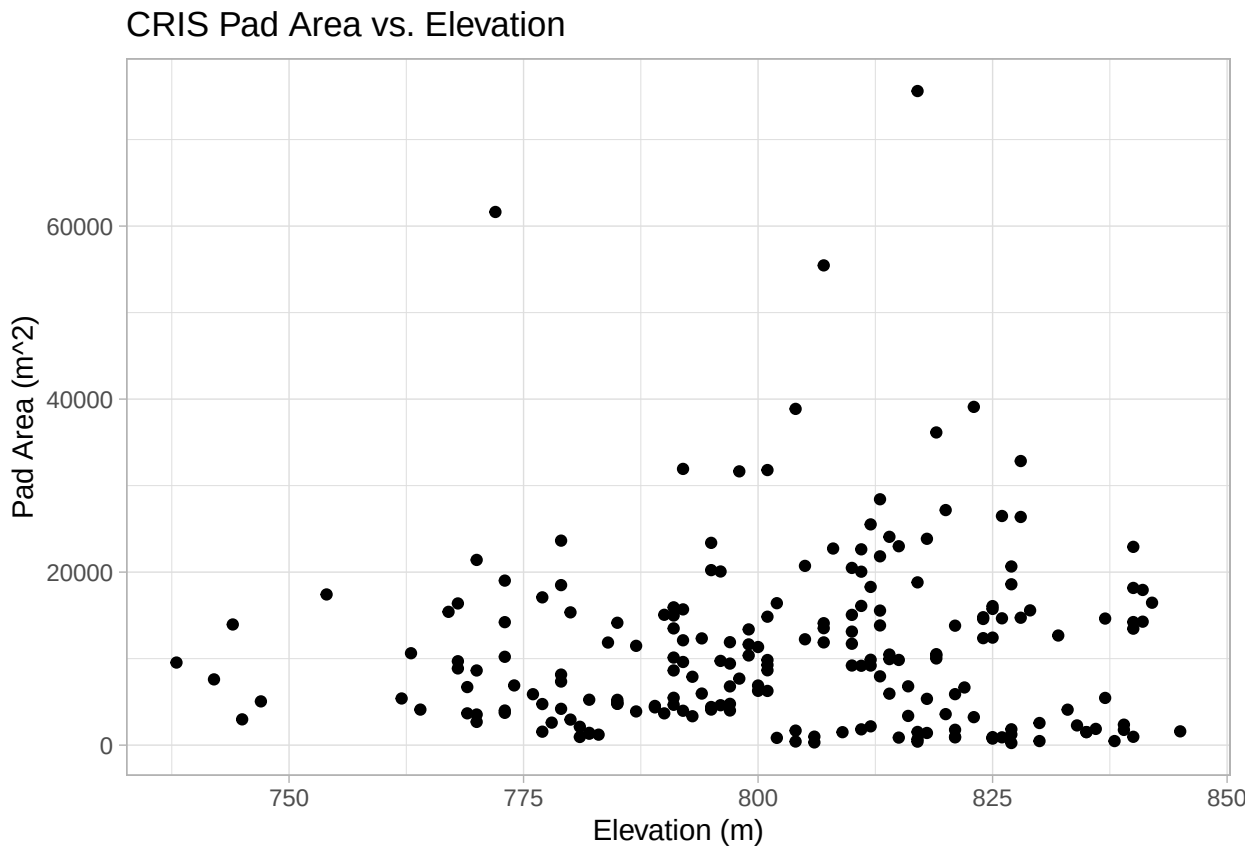
Questions

1. Read in the CRIS_AI_elev.csv file into RStudio directly from the URL address provided above and assign it to the variable "CRIS". In a single figure, show the relationship between pad area (dependent variable) and elevation. As usual, all graphics created in this assignment should be done using the ggplot2 R package and will be graded based on the highest standards of data visualization.

```
path = "http://people.tamu.edu/~geoallen/courses/312/CRIS_AI_elev.csv"
CRIS = read.csv(path, header=T)
library(ggplot2)
colorGood <- c("#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7", "#999999")
#colorGood is a set of colors suited for visibility to various types of colorblindness


ggplot() +
  geom_point(data=CRIS, aes(elev_m, y=area_m2)) +
  theme_light() +
  scale_fill_manual(values = colorGood) +
  labs(x = "Elevation (m)",
```

```
       y = "Pad Area (m^2)",
       title = "CRIS Pad Area vs. Elevation")
```



The general pattern seems to indicate/suggest (without the obvious high outliers included) that the lower bound of pad size doesn't change much regardless of elevation, but the upper bound does increase roughly linearly up to ~825m, tapering off somewhat after that.

2. Are aridity and pad area positively or negatively correlated? State the null hypothesis of correlation. Is the Pearson correlation coefficient significant at the 95% confidence interval?

```
cor.test(x=CRIS$aridity, y=CRIS$area_m2, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  CRIS$aridity and CRIS$area_m2
## t = 0.11494, df = 205, p-value = 0.9086
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1284828  0.1442399
## sample estimates:
##         cor
## 0.008027827
```

Aridity and pad area are correlated positively, albeit by a very small amount. The null hypothesis of correlation is that the correlation coefficient of these two varaibles is not meaningfully, statistically different than 0. To condense, NULL => Correlation ~= 0. As the p-value is not less than the critical value of 0.05

for the 95% confidence interval, we can reasonably assume that the coefficient is not statistically significant. Thus, we fail to reject the null hypothesis that these variables are not meaningfully correlated.

3. Are elevation and aridity positively or negatively correlated? Using the Pearson correlation coefficient, state whether this correlation is significant at the 95% confidence interval. In one sentence answer the following: does aridity cause a change in elevation, why or why not?

```
cor.test(x=CRIS$elev_m, y=CRIS$aridity, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  CRIS$elev_m and CRIS$aridity
## t = -9.4961, df = 205, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6407891 -0.4502881
## sample estimates:
##       cor
## -0.552718
```

Aridity and elevation are correlated negatively. As the p-value is substantially less than the critical value of 0.05 for the 95% confidence interval, we can reasonably assume that the coefficient is statistically significant. Thus, we can reasonably reject the null hypothesis that these variables are not meaningfully correlated. In one sentence, we can conclude that there is near-certainly an inverse relation between aridity and elevation, but we cannot be sure which is the controlling variable, if it is just one of the two tested, both, or none (i.e. correlation by coincidence).

4. Using Spearman's rank correlation coefficient, determine whether there is a statistically significant correlation between pad area and aridity. In one sentence, what about the data might explain the differences between the results of the Pearson and Spearman tests?

```
cor.test(x=CRIS$area_m2, y=CRIS$aridity, method="spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  CRIS$area_m2 and CRIS$aridity
## S = 1264360, p-value = 0.03759
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1446948
```

```
As the p-value is  less than the critical value of 0.05 for the 95% confidence interval, we can reasonal
Looking at the data, we can clearly see that pad area and aridity are nowhere near linearly correlated,
```

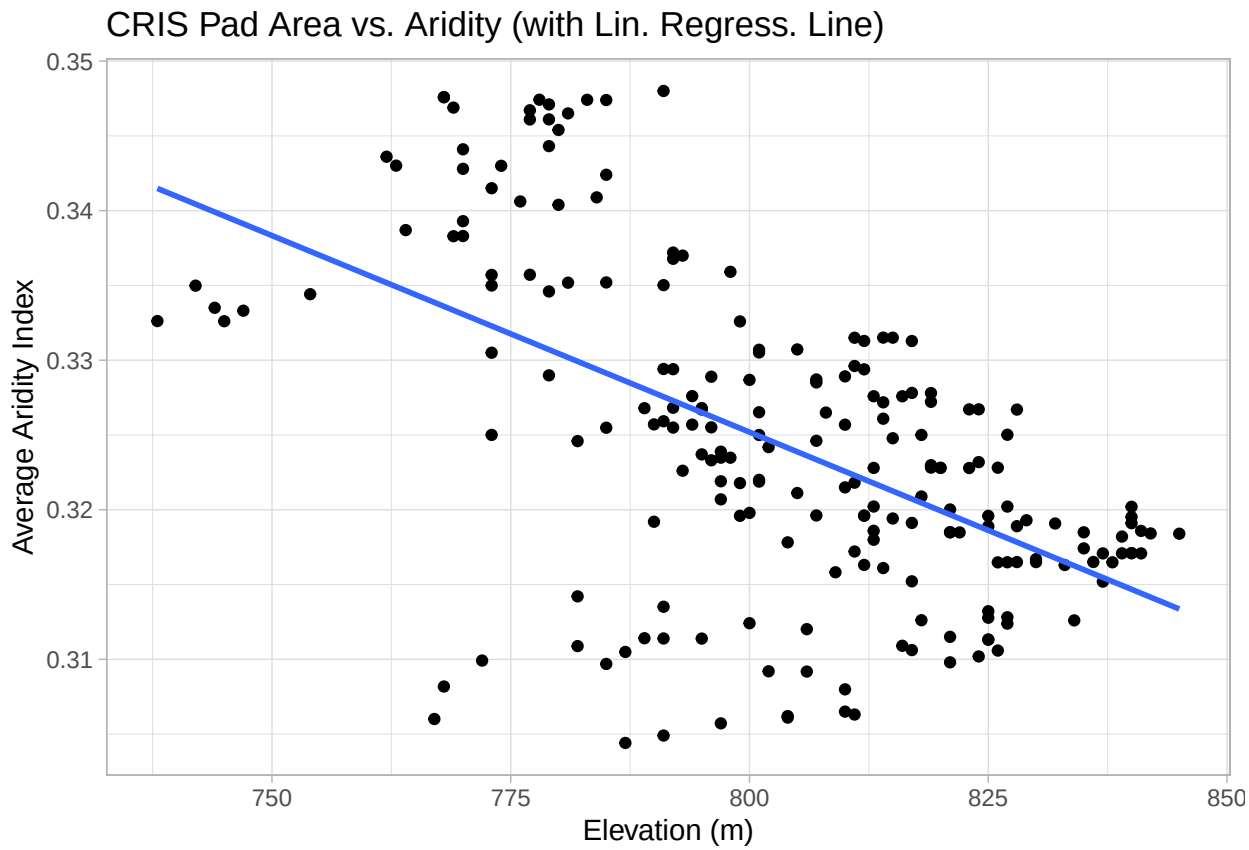
5. Perform a least-squares linear regression between elevation (independent variable) and aridity. State the values of the slope and y-intercept of this regression. Plot elevation and aridity data with ggplot2 and add the least-squares regression as a blue line to the plot. Are these data are a sample or the population of elevation and aridity in the study area?

```
lr = lm(aridity~elev_m, data=CRIS)
summary(lr)
```

```
##
```

```
## Call:
## lm(formula = aridity ~ elev_m, data = CRIS)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0278543 -0.0051871  0.0007776  0.0056288  0.0204441
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.354e-01  2.223e-02  24.087   <2e-16 ***
## elev_m      -2.628e-04  2.767e-05  -9.496   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.008994 on 205 degrees of freedom
## Multiple R-squared:  0.3055, Adjusted R-squared:  0.3021
## F-statistic: 90.18 on 1 and 205 DF,  p-value: < 2.2e-16
```

```
ggplot() +
  geom_point(data=CRIS, aes(x=elev_m, y=aridity)) +
  geom_smooth(data=CRIS, aes(x=elev_m, y=aridity), formula=y~x, method="lm", se=F) +
  theme_light() +
  scale_fill_manual(values = colorGood) +
  labs(x = "Elevation (m)",
       y = "Average Aridity Index",
       title = "CRIS Pad Area vs. Aridity (with Lin. Regress. Line)")
```



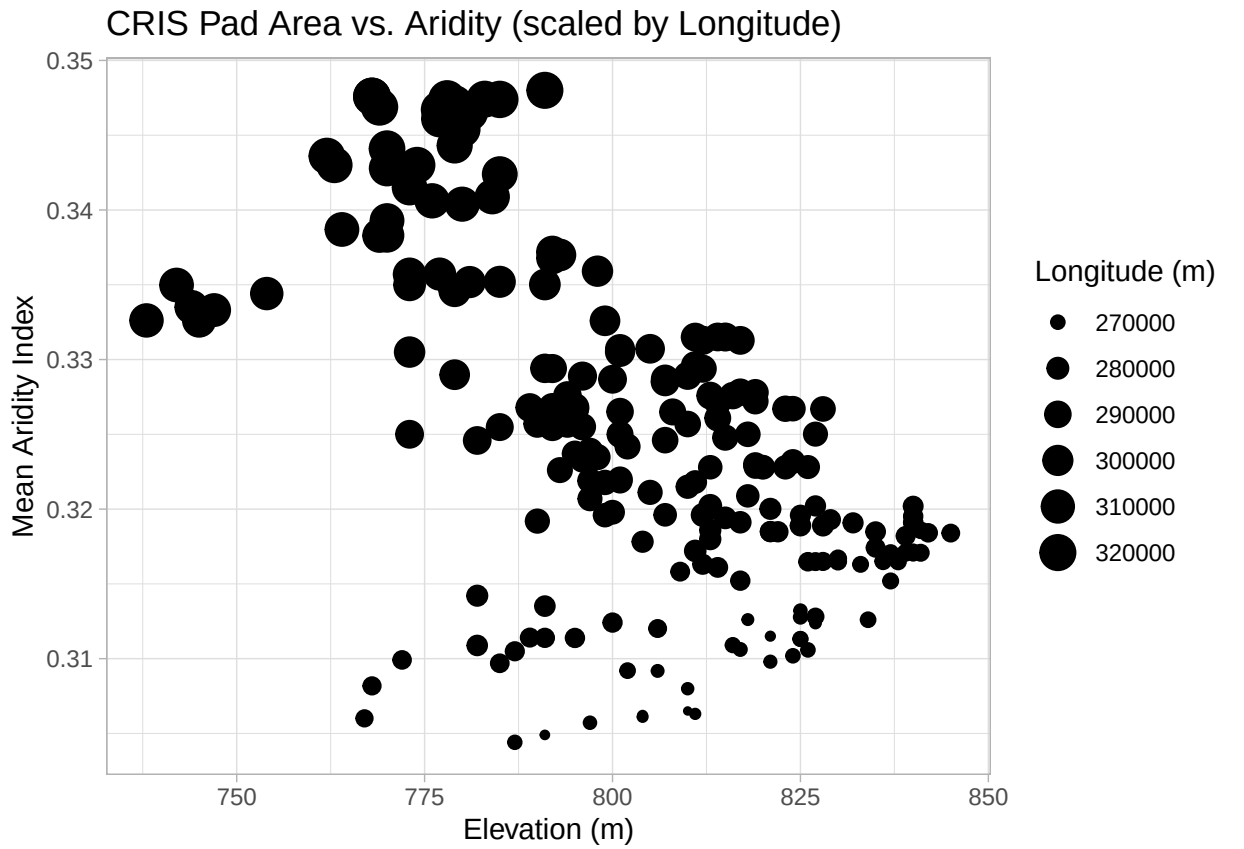
The intercept, as labeled above, is 0.5354308, and the y-intercept is -0.0002628. These data are a sample of the total population, as there are not measurements for every single pad in the CRIS Survey area.

6. Apply a multiple linear regression between aridity (dependent variable), elevation, and longitude. Plot aridity vs. elevation and scale the symbols by longitude. In a sentence, does incorporating longitude improve the ability of the model to represent the variability in aridity? How do you know?

```
mlr = lm(aridity~elev_m+x, data=CRIS)
summary(mlr)
```

```
##
## Call:
## lm(formula = aridity ~ elev_m + x, data = CRIS)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.600e-03 -5.908e-04  8.591e-05  7.848e-04  2.362e-03
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.848e-02  5.204e-03  -7.395  3.6e-12 ***
## elev_m       1.533e-04  4.549e-06  33.700  < 2e-16 ***
## x            8.325e-07  6.611e-09 125.924  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001016 on 204 degrees of freedom
## Multiple R-squared:  0.9912, Adjusted R-squared:  0.9911
## F-statistic: 1.146e+04 on 2 and 204 DF,  p-value: < 2.2e-16
```

```
ggplot() +
  geom_point(data=CRIS, aes(x=elev_m, y=aridity, size=x)) +
  theme_light() +
  scale_fill_manual(values = colorGood) +
  labs(x = "Elevation (m)",
       y = "Mean Aridity Index",
       size = "Longitude (m)" ,
       title = "CRIS Pad Area vs. Aridity (scaled by Longitude)")
```

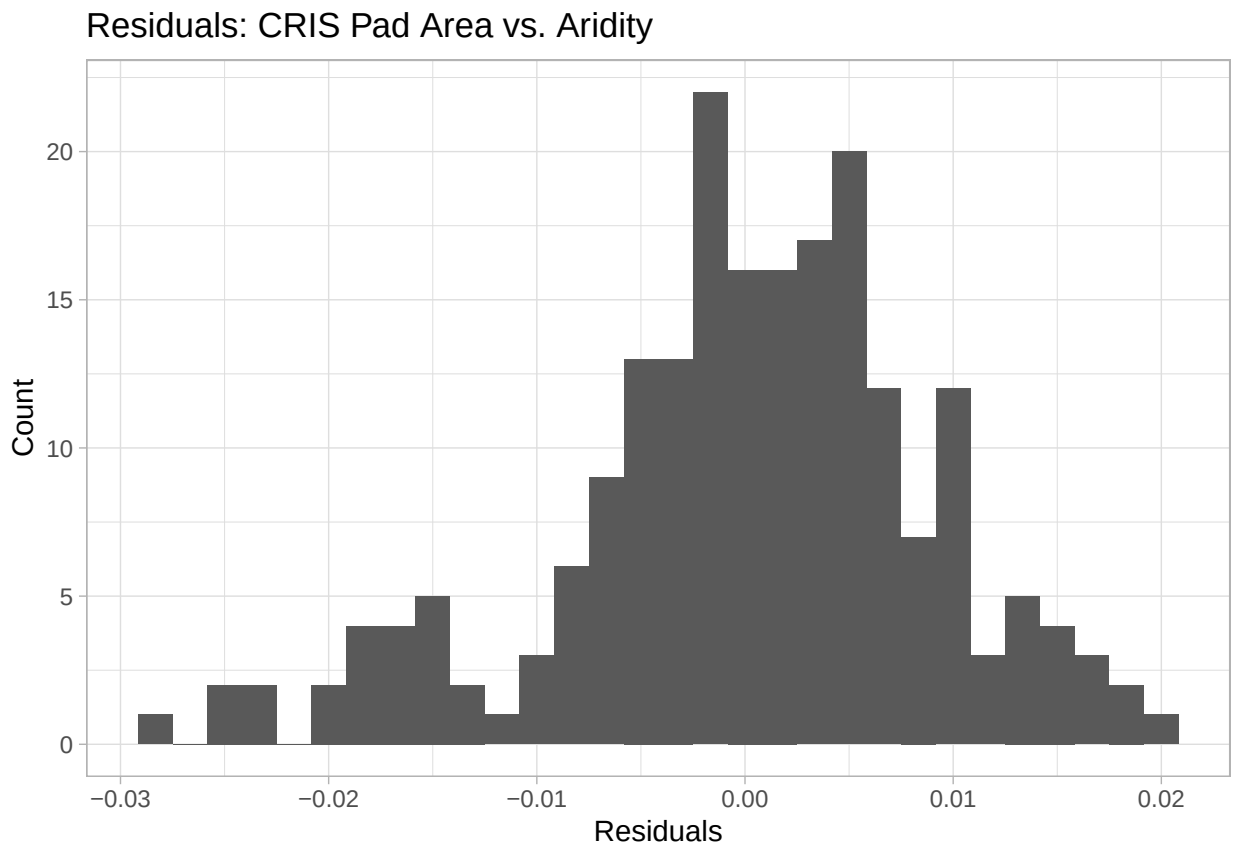


With the new information derived from the size-scaled plot, we can see that there is a close positive relation between aridity and longitude, as the longitude visibly grows along with aridity index. As both the independent variables for this question have a close relation to aridity, using both is very likely to create a model that more closely fits reality.

7. Plot the distribution of the regression residuals of Question #5's linear regression as a histogram. Calculate and print the standard error of the estimator. Are the residuals normally distributed? Hint: you could use the Shapiro-Wilk normality test. Is least squares the optimal regression approach to use in this case?

```
ggplot() +
  geom_histogram(data=lr, aes(x=lr$residuals)) +
  theme_light() +
  scale_fill_manual(values = colorGood) +
  labs(x = "Residuals",
       y = "Count",
       title = "Residuals: CRIS Pad Area vs. Aridity")
```

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




```
print( sd(lr$residuals) )
```

```
## [1] 0.008972038
```

```
shapiro.test(lr$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lr$residuals
## W = 0.97475, p-value = 0.0008851
```

```
As the p-value for the Shapiro-Wilk Test is less than the default of 0.05, we can reject the hypothesis
```

8. (Bonus 5 pts) Apply a nonparametric linear regression to aridity and elevation. Print the coefficients of the nonparametric regression. Create a new ggplot figure, composed of the same figure generated in Question #5 but with the addition of the nonparametric regression line, colored red. Which line do you think fits the data better and what about the type of regression makes it fit the data better?

```
np = loess(aridity~elev_m, data=CRIS)
summary(np)
```

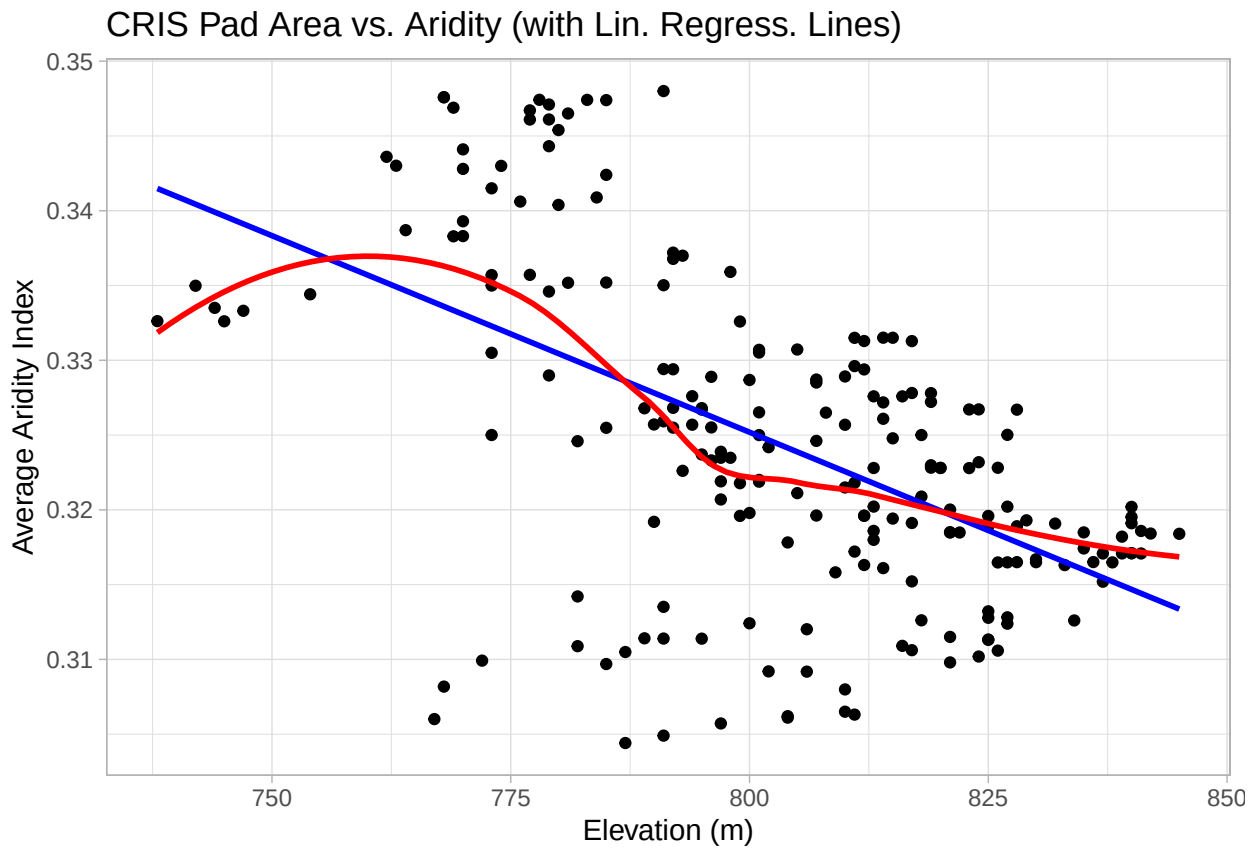
```
## Call:
## loess(formula = aridity ~ elev_m, data = CRIS)
##
## Number of Observations: 207
## Equivalent Number of Parameters: 4.95
## Residual Standard Error: 0.008711
```

```
## Trace of smoother matrix: 5.41  (exact)
##
## Control settings:
##   span     :  0.75
##   degree   :  2
##   family   :  gaussian
##   surface  :  interpolate      cell = 0.2
##   normalize:  TRUE
##  parametric:  FALSE
## drop.square:  FALSE
```

```
ggplot() +
  geom_point(data=CRIS, aes(x=elev_m, y=aridity)) +
  geom_smooth(data=CRIS, aes(x=elev_m, y=aridity), formula=y~x, method="lm", se=F, color = "blue", name=
  geom_smooth(data=CRIS, aes(x=elev_m, y=aridity), formula=y~x, method="loess", se=F, color = "red", na
  theme_light() +
  scale_fill_manual(values = colorGood) +
  labs(x = "Elevation (m)",
       y = "Average Aridity Index",
       name = "Regression Lines",
       title = "CRIS Pad Area vs. Aridity (with Lin. Regress. Lines)")
```

```
## Warning: Ignoring unknown parameters: name
```

```
## Warning: Ignoring unknown parameters: name
```



The loess function appears to create a better fit for the data, as although the line itself is not easily quantifiable (see the somewhat-jagged line span near Elevation=800m), it is more fitted towards the

individual points vs. having to do its best to model the least squares for the point distribution as a whole.