HW7

Matt Kaye and Ben Schwartz

10/28/2018

Problem:

Of the nearly 6,500 languages spoken around the world, nearly 2,000 languages are spoken by 1,000 people or less. There is typically a higher number of languages spoken in equatorial countries. In this study, we look at the hypothesis that language diversity is partially related to food security. In areas where food is more secure, groups do not rely on outside help for food, leading to more productivity within smaller groups, and leading to higher language diversity. When food is less secure, groups interact more consistently due to dependence for food, pushing social forces that decrease language diversification.

In particular, we will be examining the hypotheses: A country's language diversity is positively associated to the length of its growing season. Language diversity is negatively associated to the variability of its growing season. The average length of the growing season and variability of the growing season interact to reduce language diversity.

Data

The data for this project were collected from 74 countries located between the Tropic of Cancer and the Tropic of Capricorn. The countries considered were all relatively close to the equator. These countries often have warmer climates, which leads to longer growing seasons as well as more areas with subsistence farming, which was referenced to increase language diversity. One main reason for looking past the northern and southern countries is due to the industrial revolution which took place across Europe, Asia, and America, which in turn led to a standardization of a national language.

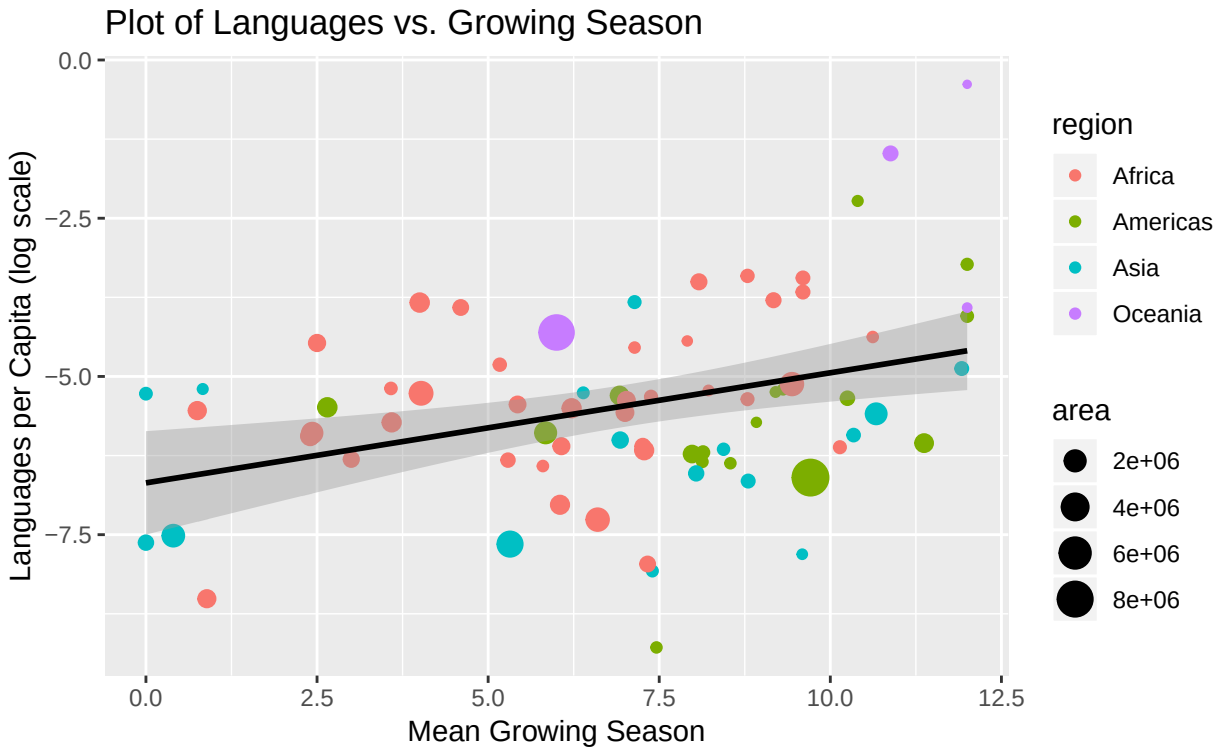
The data from the 74 countries include the length of the mean growing season, in months, constructed from a formula from Le Houe´rou 1989, which looks at average temperature and rainfall. There is also data on the area of each country, in km^2, the population (in 1000s), as well as the standard deviation of the growing season.

Additionally, we assigned each country to a region (i.e. Africa, Americas, Asia, and Oceania) to help with our EDA, and to help us understand how the relationships that we observe vary by region.

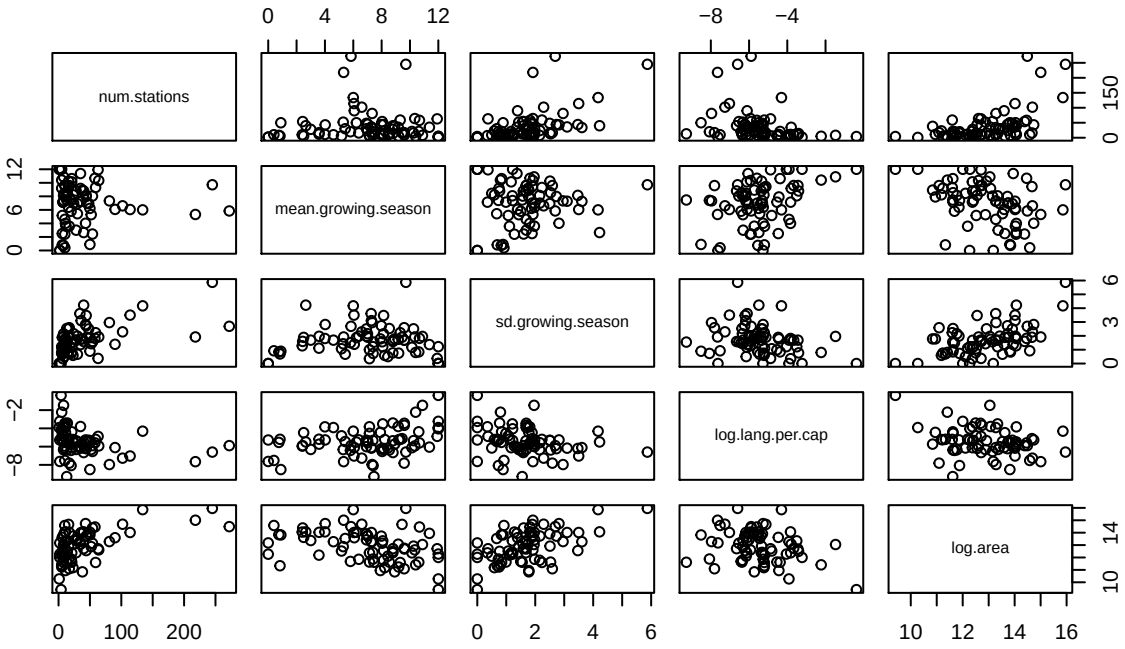
To get a sense of the data, we performed some basic EDA:

EDA

We plot log of the number of languages spoken per capita against the mean growing season here. The colors represent different regions, and the bubble sizes represent the populations of the countries.



Here we make a scatter plot matrix of our quantitative variables of interest.



Methods:

In general, as we are trying to infer about the association between language diversity and the length and diversity of the growing season, a linear model will provide the most easily interpretable results about these relationships. If we were trying to make predictions about out-of-sample points, a higher order model may be useful to generate better out of sample predictions, but not in the context of these hypotheses. For these models, we used diffuse priors so that the data may overcome the priors, giving us estimates which allow the data to influence our results more than our specified priors.

The first hypothesis, which states that the length of a country's growing season is positively associated with its language diversity is best tested using multiple regression. A linear relationship between the growing season and language diversity allows us to, while holding other variables constant, make clear inferences in how a change in mean growing season will affect the number of languages per capita, on the log scale.

$$Y = languages\ per\ capita$$

$$log(Y) = \alpha + \beta_1 * mean.growing.season + \beta_2 * log(area)$$

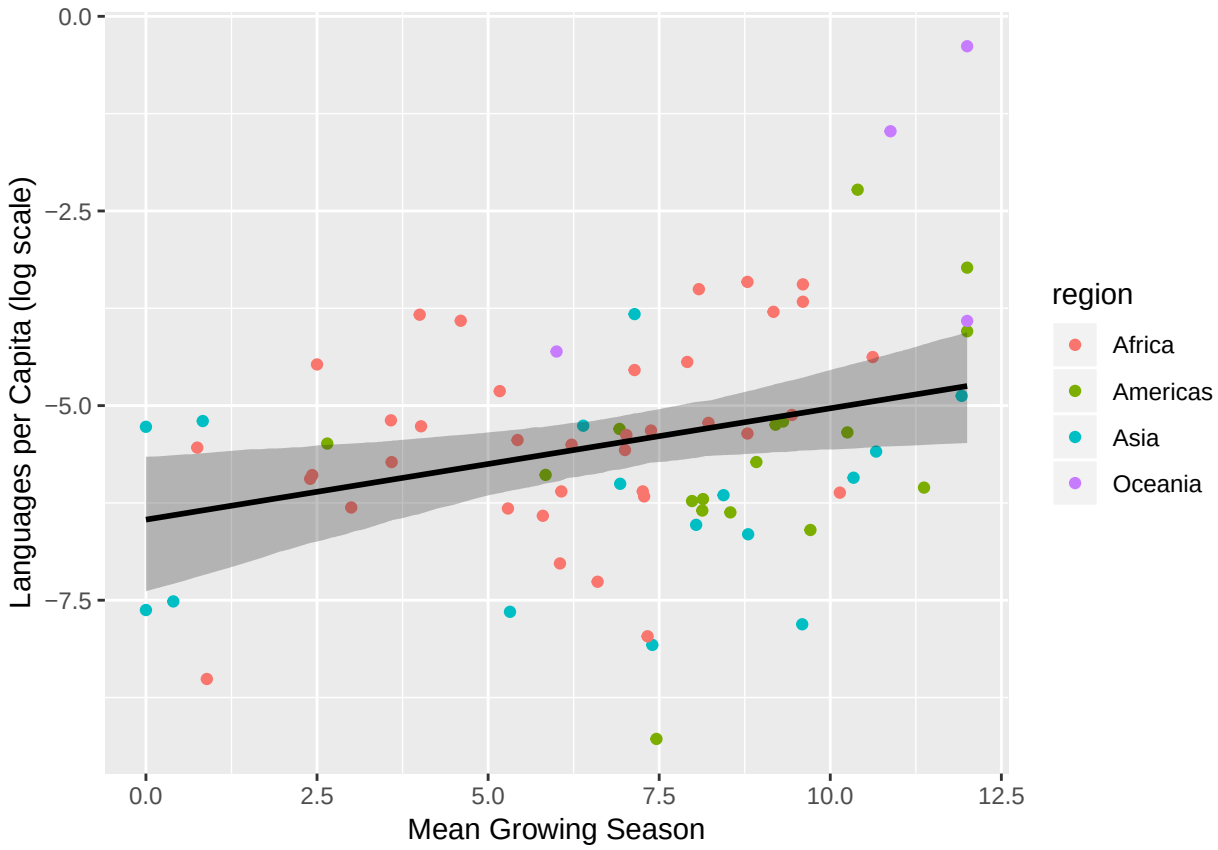
We use a linear model to test the effect of the mean length of the growing season in a country on the log of its number of languages spoken. From our model, we can see that there is a significant positive association between mean growing season and language diversity because the β_1 is statistically significant. Thus, we can reject the null hypothesis that there is no effect and assert that the average length of the growing season positively affects the log of the number of languages spoken in a country.

Below are the results of our linear model.

Table 1: Model 1 Output

	Mean	StdDev	5.5%	94.5%
Intercept	-3.854	1.962	-6.990	-0.718
Mean Growing Season	0.144	0.056	0.055	0.233
log(Area)	-0.202	0.137	-0.422	0.017
sigma	1.389	0.114	1.207	1.572

We plot the log of the number of languages per capita against the mean growing season. As above, color represents geographic region. In this plot, the line is the regression line that we get from fitting our model, and the shaded bands represent a 97% prediction interval of the mean.



Our second hypothesis, that the variability of the growing season being negatively associated with the number of languages per capita, is also best tested using multiple regression. In this case, instead of mean growing season we use the standard deviation of a country's growing season.

We constructed a multiple regression model similar to hypothesis 1 to draw inferences from our data.

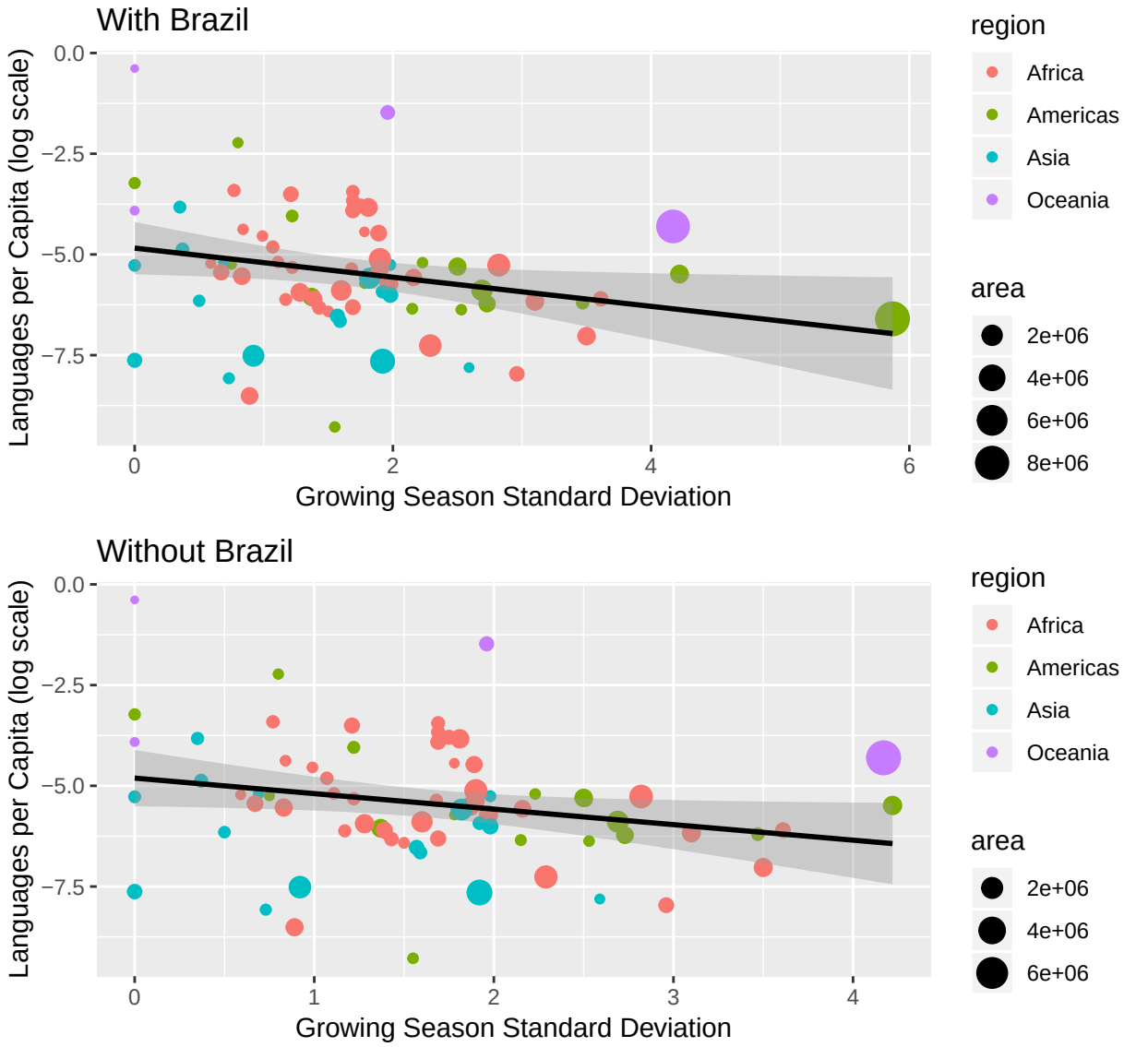
$$log(Y) = \alpha + \beta_1 * sd.growing.season + \beta_2 * log(area)$$

Below are the results of our linear model.

Table 2: Model 2 Output

	Mean	StdDev	5.5%	94.5%
Intercept	-1.998	1.878	-4.999	1.002
sd Growing Season	-0.209	0.186	-0.507	0.089
log(Area)	-0.240	0.156	-0.489	0.010
sigma	1.438	0.118	1.249	1.627

Something to note here is that Brazil shows up as an outlier in our model. See the following plots:



In this model we test the hypothesis that the standard deviation of the length of the growing season is statistically significant in predicting the log of the number of languages spoken per capita in each country. We find that the β_1 is insignificantly different than zero, which means that the standard deviation of the length of the growing season is not significant in predicting the log of the number of languages spoken on a per capita basis.

The plots included here show the relationship between these two variables with and without Brazil, which looks like an outlier. The different colors show different regions of the world, and the sizes of the bubbles represent the land areas of the countries. Our original model included Brazil and we could not reject the null that the true relationship was not zero, and after dropping Brazil the slope looks even flatter, indicating that the true relationship may possible be the null hypothesis.

To test the hypothesis that the average length of the growing season and the standard deviation of the length of the growing season in a country act together in predicting the number of languages spoken, we add an interaction term between the two predictors to our model. The idea is that if the interaction term is significant, then the theory that suggests an important relationship between the mean and standard deviation of the length of the growing season that is outlined in the problem holds.

We tested two models in addressing this problem: one including the log of the area and one not. The

comparison of the two models is below, along with the results of the model without the area term.

Table 3: Model Comparison (With and Without Area Term)

	WAIC	pWAIC	dWAIC	weight	SE	dSE
mod.c.without	261.148	5.764	0.000	0.749	15.876	NA
mod.c.with	263.339	6.801	2.191	0.251	16.086	0.489

Table 4: Model 3 Output

	Mean	StdDev	5.5%	94.5%
Intercept	-6.972	0.591	-7.917	-6.027
sd Growing Season	0.421	0.372	-0.174	1.016
Mean Growing Season	0.299	0.072	0.184	0.415
Interaction	-0.109	0.047	-0.184	-0.034
sigma	1.306	0.107	1.135	1.478

Since the log of area was insignificant in predicting the log of the number of languages spoken and the WAIC of the model that did not include the area term was smaller than the one that did, we decided to use the model without the term for area. In other words, we drop the area term because it does not seem to be contributing to the model in any significant way.

Our linear model is the following:

$$M = mean.growing.season$$

$$S = sd.growing.season$$

$$log(Y) = \alpha + \beta_1 \cdot M + \beta_2 \cdot S + \beta_3 \cdot M \cdot S + \beta_4 * log(area)$$

In our interaction model, we noticed that β_4 was centered nearly at zero. Because we thought this term might only clutter our model, we removed it as a covariate term and re-estimated our model parameters. After removing this term, we compared the WAIC estimates for each model, with and without the area term, and concluded that the WAIC estimate for the model without area is better than with area. As a result we used the model without area as a covariate term to infer about our interaction term between the length and standard deviation of the growing season.

In our results, we find that the interaction term is statistically significant, so we reject the null hypothesis that there is no important relationship between the mean and standard deviation of the length of the growing season in favor of the theory that the relationship between the two has an effect on the importance of storage and redistribution of crops, which leads to more cooperation between groups and less languages spoken.

In order to draw conclusions, we looked at the three hypotheses:

1. H_A: The length of a country's growing season is positively associated with its language diversity.

2. H_A: The standard deviation of the growing season being negatively associated with the log number of languages per capita.

3. H_A: The average length of the growing season and the standard deviation of the length of the growing season in a country act together and are negatively associated to affect the number of languages spoken.

Through multiple regression, we found that the hypotheses (1) and (3) were significant and we can accept those hypotheses. Therefore, we can conclude that countries with longer growing seasons will have a higher amount of languages per capita, on the log scale, holding all else constant. We also see that the variability of the length of the growing season interacts with the average length of the growing season, decreasing the number of languages per capita, on the log scale.

In our second model, we found that we cannot reject the null hypothesis, meaning that we cannot accept our proposed hypothesis. This is interesting considering that more variable climates might not have any effect on the number of languages within a given country. This is an interesting conclusion when looking at the practical meaning of this. One could initially think that countries with more variable growing seasons will rely more on outside groups for food, decreasing language diversity due to interactions with other groups of people. Our model tells us that this interaction is possibly insignificant, which is almost counterintuitive.

Overall, looking at our models in the context of the problem provides us with some interesting insights. We see that in places with longer growing seasons, we expect more languages per capita. As Nettle referenced in his study, the longer growing seasons allow for individual groups to be able to provide food internally. This, in turn, means that these people will not need to rely on other groups for their food sources, which prevents the mingling of languages. As a result, we remain with more unique languages, and a higher language diversity.

Going forward, it would be interesting to test these results again with more current data, to see if our inferences have changed over the past twenty years or so.