Xuan Hao Xu
Winter 2015
VEE Regression

Temperature, Precipitation, and Life Expectancy

Introduction

We constantly hear about factors that affect our health, which can include air quality, food quality, living quality, etc. There are many factors that contribute to our life expectancy. Is geographic location one of them? As we know, the majority of a population in one specific geographic area has adapted to the natural environment there after thousands of years. In theory, life expectancy should be constant among various geographic locations, but this is not necessarily true. Temperature and Precipitation are two characteristics of a geographic area. I would like to explore how these two factors affect life expectancy.

In the process of this analysis, I will be comparing four models incorporating the two explanatory variables, Temperature and Precipitation, to attempt to explain the response variable Life Expectancy. At the end, I will choose the best model of the four by analyzing the graphs and performing F-tests.

Data

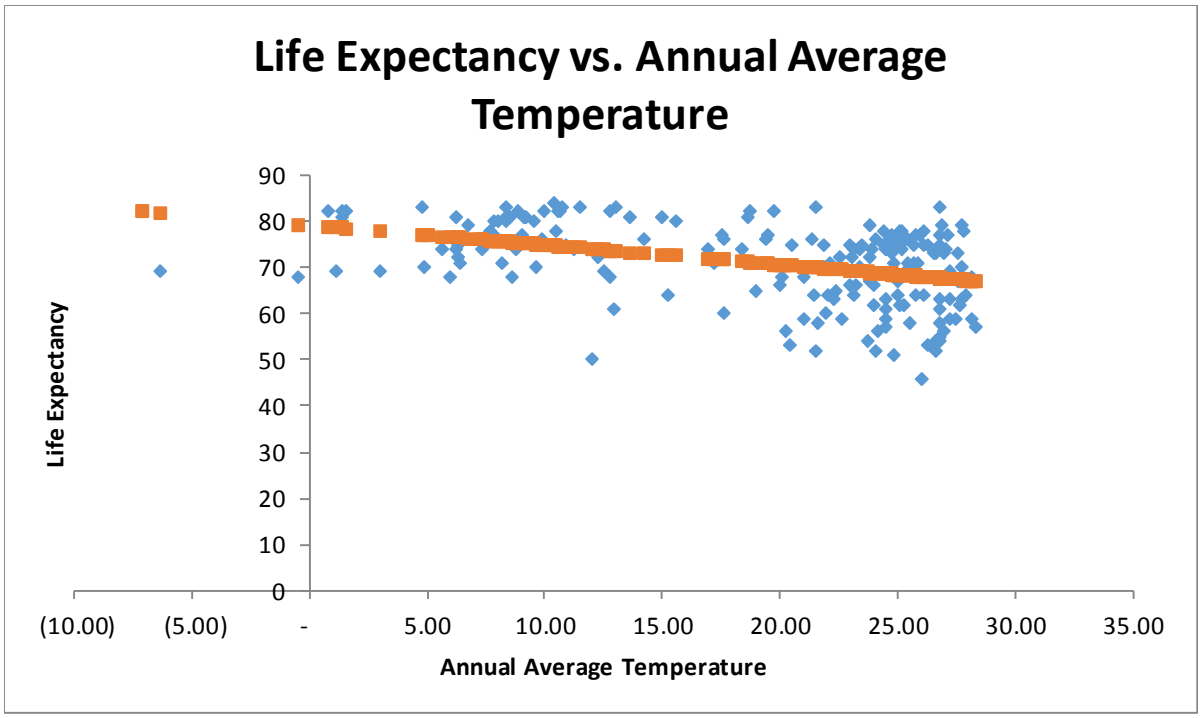
Please refer to the attached Excel worksheet for data. The data includes Life Expectancy by country (which defines a geographic area) at birth (2013) published by the World Health Organization in 2015, and average annual temperature and annual precipitation by country using years 1961 to 1999 published by the World Bank.

We start by looking at whether Temperature and Precipitation are correlated. I found that the correlation between temperature and precipitation is 0.4069. This shows that temperature and precipitation are positively correlated. When producing the full model, I will be adding an interaction term to explain this correlation.

Model 1: How does temperature affect life expectancy?

$$yi = alpha + beta \times Tempi + \varepsilon i$$

First, we will look at how the average temperature alone affects life expectancy.



From the graph, we can see a negative relationship between annual average temperatures: temperature increases, life expectancy decreases. There also seems to be two groups of life expectancy, temperature between 5 – 10 degrees Celsius and 20 – 30 degrees Celsius. There are only a few countries with average temperatures between 10 and 20 degrees Celsius.

The regression statistics for this model is shown below:

Regression Statistics	
Multiple F	0.42
R Square	0.17
Adjusted I	0.17
Standard I	7.82
Observati	194.00

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressio	1.00	2,451.16	2,451.16	40.09	0.00
Residual	192.00	11,738.10	61.14		
Total	193.00	14,189.26			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	79.08	1.39	57.07	0.00	76.35	81.82	76.35	81.82
Annual_te	(0.43)	0.07	(6.33)	0.00	(0.56)	(0.29)	(0.56)	(0.29)

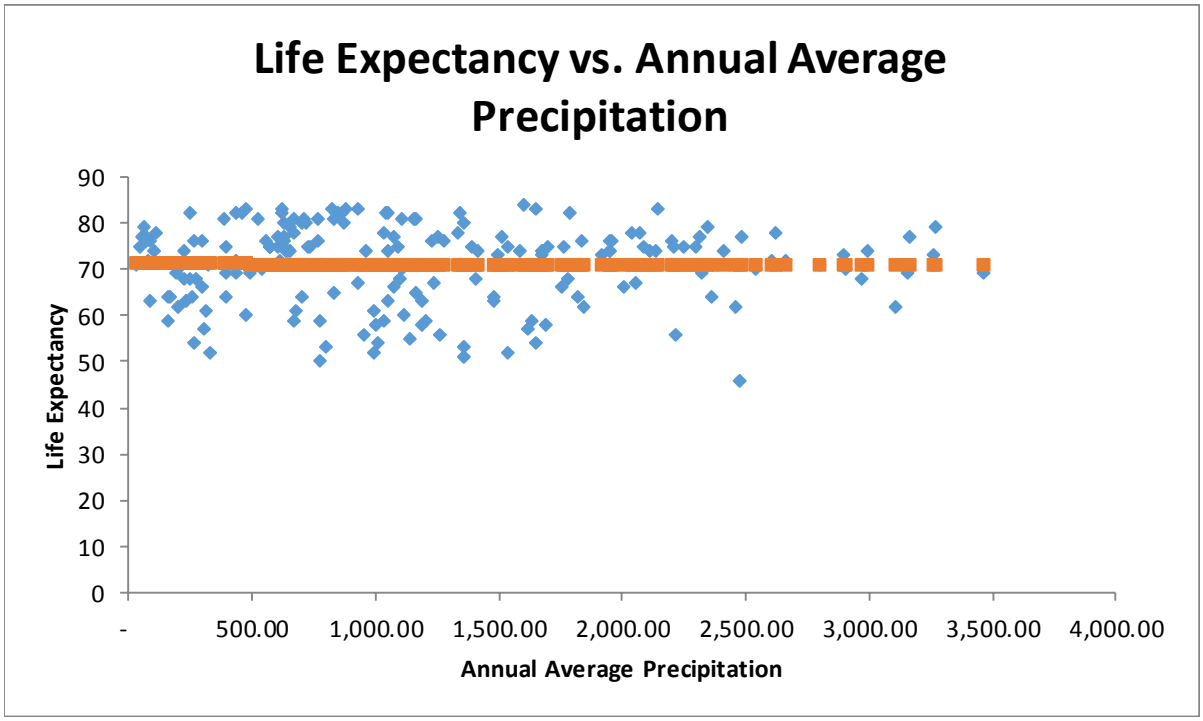
$$y_i = 79.08 - 0.43 \times Temp_i + \varepsilon_i$$

Correlation coefficient R = -0.42, which indicates a negative moderate correlation between Temperature and Life Expectancy.

Model 2: How does precipitation affect life expectancy?

$$y_i = alpha + beta \times Prep_i + \varepsilon_i$$

We will look at the second model with precipitation being the only explanatory variable:



From the graph, it seems that the linear regression has 0 slope and nothing seems to be explained by the model.

Regression Statistics	
Multiple	0.01
R Square	0.00
Adjusted	(0.01)
Standard	8.60
Observat	194.00

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressic	1.00	1.83	1.83	0.02	0.87
Residual	192.00	14,187.42	73.89		
Total	193.00	14,189.26			

	Coefficients	*tandard Errc*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	71.20	1.09	65.49	0.00	69.06	73.35	69.06	73.35
Annual_p	(0.00)	0.00	(0.16)	0.87	(0.00)	0.00	(0.00)	0.00

$$y_i = 71.20 - 0.0001 \times Temp_i + \varepsilon_i$$

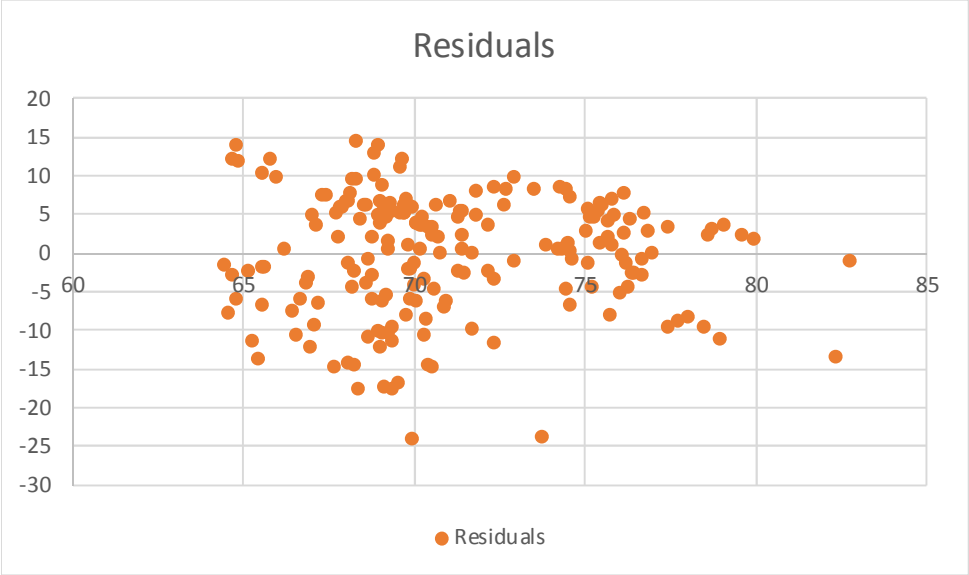
From the ANOVA table above, we can see that there is a slightly negative slope indicating that annual precipitation also has a negative relationship with life expectancy. Since the slope of this regression model is so small and 0% of the data is explained by this model (indicated by the coefficient of determination R^2), they led me to think whether precipitation is a significant explanatory variable.

Correlation coefficient R = -0.01, which indicates poor/no correlation between Precipitation and Life Expectancy.

Model 3: Precipitation and Temperature are explanatory variables.

$$y_i = alpha + beta_{1i} \times Temp_i + beta_{2i} \times Prep_i + \varepsilon_i$$

Setting Temperature and precipitation both as explanatory variables for life expectancy, the residual graph is shown below with fitted y values.



The residuals shown above is quite random and centered around 0. It shows that we have a good fit of the response variable.

Regression Statistics	
Multiple	0.45
R Square	0.20
Adjusted	0.19
Standard	7.70
Observat	194.00

R = 0.4501

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressic	2.00	2,874.32	1,437.16	24.26	0.00
Residual	191.00	11,314.93	59.24		
Total	193.00	14,189.26			

	Coefficients	*tandard Errc*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	78.22	1.40	55.81	0.00	75.46	80.99	75.46	80.99
Annual_t	(0.51)	0.07	(6.96)	0.00	(0.65)	(0.36)	(0.65)	(0.36)
Annual_p	0.002	0.00	2.67	0.01	0.00	0.00	0.00	0.00

$$y_i = 78.22 - 0.51 \times Temp_i + 0.002 \times Prep_i + \varepsilon_i$$

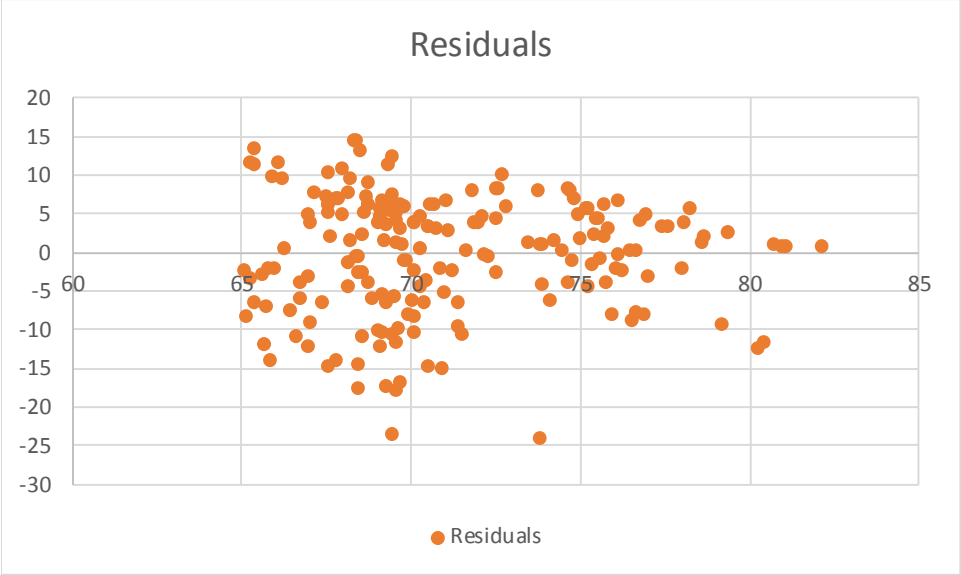
Looking at the regression statistics, the coefficient of determination (R^2) is definitely an improvement over Model 1. This indicates that 20% of the data is explained by the model compared to 17% in Model 1 and 0% in Model 2. At the end of Model 2, I mentioned that I doubt that Precipitation is a significant explanatory variable, but I might be wrong from analyzing this model. The improvement might be stimulated by the correlation between temperature and precipitation. The coefficient for temperature is negative and precipitation

positive. This result is different from Model 1 and 2 where coefficients for both temperature and precipitation are negative. I would like to examine the full model before making any conclusions.

Model 4: Full model with interactions between Temperature and Precipitation

$$y_i = alpha + beta_{1i} \times Temp_i + beta_{2i} \times Prep_i + beta_{3i} \times Temp_i \times Prep_i + \varepsilon_i$$

As identified previously, Temperature and Precipitation have a correlation of 0.41. Therefore, I am adding an interaction term to see how this could improve the modeled results.



From the graph above which shows residuals by fitted y-values, we can see that the residuals are centered around 0 and they seem to have separated into two groups: fitted y-value between 65 and 71 and between 75 and 80. Previously, I commented that there seems to be two groups of life expectancies when graphed by temperature and fewer number of countries fall into the bucket of 10 and 20 degrees Celsius for annual average temperature. As I confirmed, the reason for having two separate groups in this residuals graph is due to lack of countries with average temperature between 10 and 20 degrees.

Regression Statistics	
Multiple R	0.46
R Square	0.21
Adjusted R Square	0.20
Standard Error	7.68
Observations	194.00

R = 0.4575

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3.00	2,969.78	989.93	16.76	0.00
Residual	190.00	11,219.48	59.05		
Total	193.00	14,189.26			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	74.88	2.98	25.15	0.00	69.01	80.76	69.01	80.76
Annual_temp	(0.36)	0.14	(2.65)	0.01	(0.63)	(0.09)	(0.63)	(0.09)
Annual_precip	0.01	0.00	1.81	0.07	(0.00)	0.01	(0.00)	0.01
Temp * Prep	(0.00)	0.00	(1.27)	0.21	(0.00)	0.00	(0.00)	0.00

From the statistics above, we can see that this is a slight betterment of Model 4 by adding in an interaction term between precipitation and temperature. As calculated before, we know that temperature and precipitation are correlated. The interaction term Temp * Prep being negative somewhat canceled out the positive slope on precipitation. Also the improvement (R^2 = 0.20 vs 0.21) over Model 3 is slight. Is the interaction term Temperature * Precipitation significant?

Comparing Models 1 – 4

The table below compares the four models analyzed above. The greater the RegSS, the higher the predictive power of the model.

Model	Variables	RegSS	df	MS	F-Test
1	Temp	2451	1	2451	40.09
2	Prep	2	1	2	0.02
3	Temp, Prep	2874		1437	24.26
4	Temp, Prep, Temp x Prep	2970	3	990	16.76

	SS	df
TSS	14189	193
RSS	11219	190

We can see that Model 4, the full model, has the highest RegSS better predictive power on life expectancy than other models. Model 2 with Precipitation as the only explanatory variable has the lowest RegSS of 2, which indicates it's not very predictive of life expectancy.

Test 1: H_0: Prep = 0

I performed an F-test using Model 3 and Model 1 to test if Precipitation is significant.

F-Test = (2874 – 2451)/(2-1)/11219*190 = 7.17

The critical value of $F^{-1}(0.95, 1, 190) = 3.891$. F-test value = 7.17 > 3.891 indicating that the F-value falls into the critical region of the F-distribution. Therefore, we should reject the null hypothesis.

Test 2: H_0: Temp = 0

I performed an F-test using Model 3 and Model 2 to test if Temperature is significant.

F-Test = (2874 – 2451)/(2-1)/11219*190 = 48.65

The critical value of $F^{-1}(0.95, 1, 190) = 3.891$. F-test value = 48.65 > 3.891 indicating that the F-value falls into the critical region of the F-distribution. Therefore, we should reject the null hypothesis.

Test 3: H_0: Temp*Prep = 0

F-Test = (2970 – 2874)/(3-2)/11219*190 = 1.62

The critical value of $F^{-1}(0.95, 1, 190) = 3.891$. F-test value = 1.62 < 3.891 indicating that the F-value is not in the critical region of the F-distribution. Therefore, we should not reject the null hypothesis.

The test results are summarized below:

Variable	Models	F Test	Critical Value	Decision
Prep	3-1	7.17	3.891	Reject
Temp	3-2	48.65	3.891	Reject
Temp * Prep	4-3	1.62	3.891	Do not Reject

Conclusion

After testing the four models described above, I can conclude that Temperature and Precipitation are significant in explaining life expectancy. Temperature alone can explain 17% of life expectancy. This is quite significant as we know that many other factors contribute to a person's life span (the other 83%), including inherent disease, culture, civil wars, ecosystem, etc.

Of the 4 models tested above, Model 3 with Temperature and Precipitation as explanatory variables in predicting the Life expectancy is the best model. First of all, temperature and precipitation are both significant explanatory variables in the model. Secondly, the null hypothesis of interaction term between temperature and precipitation equaling 0 was not rejected, which indicates Model 4 was not a significant improvement over Model 3. Lastly, the residual graph exhibiting randomness and centering around 0, both show characteristics of a good fit. To conclude, Model 3 with both Temperature and Precipitation being the explanatory variables without the interaction term is appropriate.

Therefore, we have:

$$Life\ Expectancy = 78.22 - 0.51 \times Temp_i + 0.002 \times Prep_i + \varepsilon_i.$$