Statistics Project – Math 141

Dashiell Ward, Gavin Rimmer, Bailee Brunsmann, Harrison Nicholls

CIA Factbook

Introduction

Hypothesis

There is a correlation between the proportion of the population using the internet and life expectancy at birth within a given nation.

$H_0 : B_1 = 0$

$H_A : B_1 \neq 0$

Linear model with Africa:

$y_{lifeexpectancyatbirth} = B_0 + B_1 * x_{internetusership}$

$B_0 = 63.71$ $B_1 = 23.46$

Linear model without Africa:

$y_{lifeexpectancyatbirthwithoutAfrica} = B_0 + B_1 * x_{internetusership}$

$B_0 = 69.52$ $B_1 = 13.53$

Approach, Declaration of Goals:

There is potentially a correlation between the variables. The average life expectancy in a country could influence internet usage and/or internet usage in a country could influence the average life expectancy.

Through our analysis, our goals are to observe how internet usage and life expectancy affect each other in different nations. Further exploration of this correlation, specifically in countries in Asia, is provided in the first background research article listed. Both background research articles provide insight about other factors that could be influencing or are influenced by the variables we chose to focus on. Economic development appears to have an important connection to technological advancements, which impacts internet accessibility and further, internet usage. Additional research related to this subject may be found below.

Lee, Cheng-Wen, The Relationship between Internet Environment and Life Expectancy in Asia

Article explores this question specifically in Asia, with relevant take-away in emphasising the disparity between countries with advanced telecommunication services and those without in regards to their general economic development in a globalized market.

Alzaid, Ahmed, Musleh Alsulami, Komal Komal, Adel-Maraghi, Examining the Relationship between the Internet and Life Expectancy

Article explores this question Globally, finding that the economic development of a country is greatly bolstered by internet development, and has both direct and indirect impacts on the average life expectancy of its citizens.

Data Exploration

Dataset:

Our dataset was Details on Countries, and it was sourced from the CIA factbook. A summary of variables may be found below.

Variables:

Country - Categorical - Nominal Countries recognized by the CIA.

Area - Numerical - Continuous Land in Square km

Infant Mortality Rate - Numerical - Continuous Infant mortality rate compares the number of deaths of infants under one year old in a given year per 1,000 live births in the same year. This rate is often used as an indicator of the level of health in a country.

Population - Numerical - Discrete Population compares estimates from the US Bureau of the Census based on statistics from population censuses, vital statistics registration systems, or sample surveys pertaining to the recent past and on assumptions about future trends.

Population growth rate - Numerical - Continuous Population growth rate compares the average annual percent change in populations, resulting from a surplus (or deficit) of births over deaths and the balance of migrants entering and leaving a country. The rate may be positive or negative.

Birth Rate - Numerical - Continuous Birth rate compares the average annual number of births during a year per 1,000 persons in the population at midyear; also known as crude birth rate.

Death rate - Numerical - Continuous Death rate compares the average annual number of deaths during a year per 1,000 population at midyear; also known as crude death rate.

Net migration rate - Numerical - Continuous Net Migration rate compares the difference between the number of persons entering and leaving a country during the year per 1,000 persons (based on midyear population).

Maternal mortality rate – Numerical - Continuous The Maternal mortality rate (MMR) is the annual number of female deaths per 100,000 live births from any cause related to or aggravated by pregnancy or its management (excluding accidental or incidental causes).

Life expectancy at birth – Numerical - Discrete Life expectancy at birth compares the average number of years to be lived by a group of people born in the same year, if mortality at each age remains constant in the future. Life expectancy at birth is also a measure of overall quality of life in a country and summarizes the mortality at all ages.

Internet users – Numerical - Discrete Internet users compares the number of users within a country that access the Internet. Statistics vary from country to country and may include users who access the Internet at least several times a week to those who access it only once within a period of several months.

Continent – Categorical - Nominal Continents as corresponds to each Country recognized by the CIA.

Density – Numerical - Continuous Density as the average population per square kilometer of the country. This is based on the Population and the Area data points by the CIA Factbook.

Internet Usage Proportion – Numerical - Continuous Internet Usage as the number of users within a country that access the internet. This is based on the Internet Users and the Population data points by the CIA Factbook.

> *Note: We used the "countrycode" package in R to add the categorical variable of continent to our data.*

Methods

To begin, two new columns were created in the dataframe: internet usage proportion, and continents. Internet usage proportion was calculated by dividing the total internet users in a country by its population, and would be a more effective tool in comparing countries of vastly different populations. The continents column made use of the "countrycode" package in order to further distinguish the data in the CIA Factbook geographically.

A basic scatter plot of life expectancy versus internet usage proportion was created to get a preliminary idea of the relationship. At this point it was noticed that, when graphing life expectancy against internet usage while coloring the points by continent, almost the entire lower section of outliers was comprised of countries in Africa. It was therefore theorized that internet usage only begins to correlate linearly with life expectancy at some threshold value (~65 years), which few African countries meet. Based on this, it was decided to see how the data would look with Africa excluded from the initial scatterplot.

From here, parallel operations were performed on the data set including Africa and the data set excluding Africa. Linear models were made for both and their R^2 values taken into account. As R^2 produced unexpected results (excluding Africa yielded a lower R^2, although it looked more linear), a histogram and scatter plot of the respective model's residuals were created to test the normalcy of their variation.

Given the hypothesis, a randomization test was performed to determine whether the slopes of the linear models were statistically significant. A bootstrap distribution of slopes was performed for both sets of data (those including and excluding Africa), and the quantile method was employed to calculate a confidence interval.

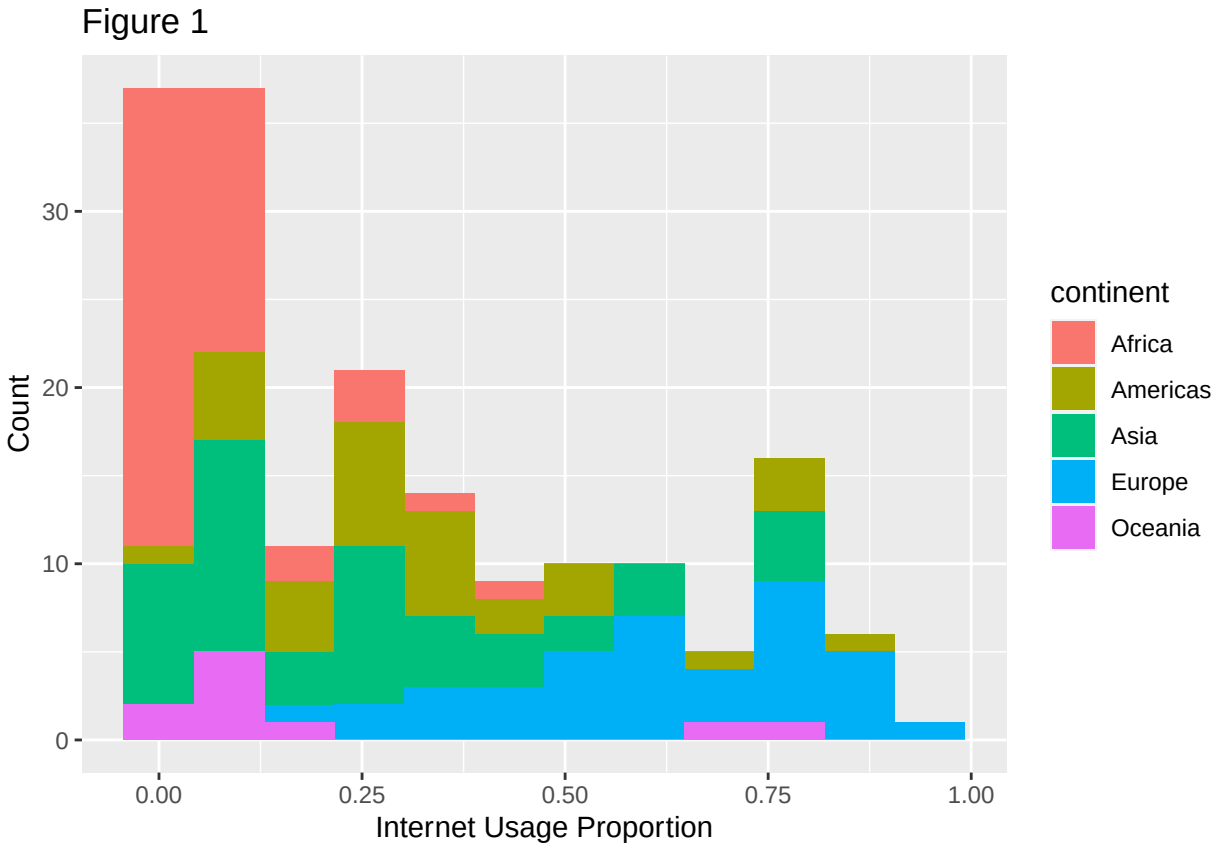


Figure 1: Most countries in Africa have an internet usage proportion less than a quarter. Asia has a leftward skew while Europe has a rightward skew. The Americas and Oceania are more evenly and sporadically distributed across the histogram. When ignoring Africa, the graph follows a very shallow negative slope, but when including Africa the slope is much steeper.

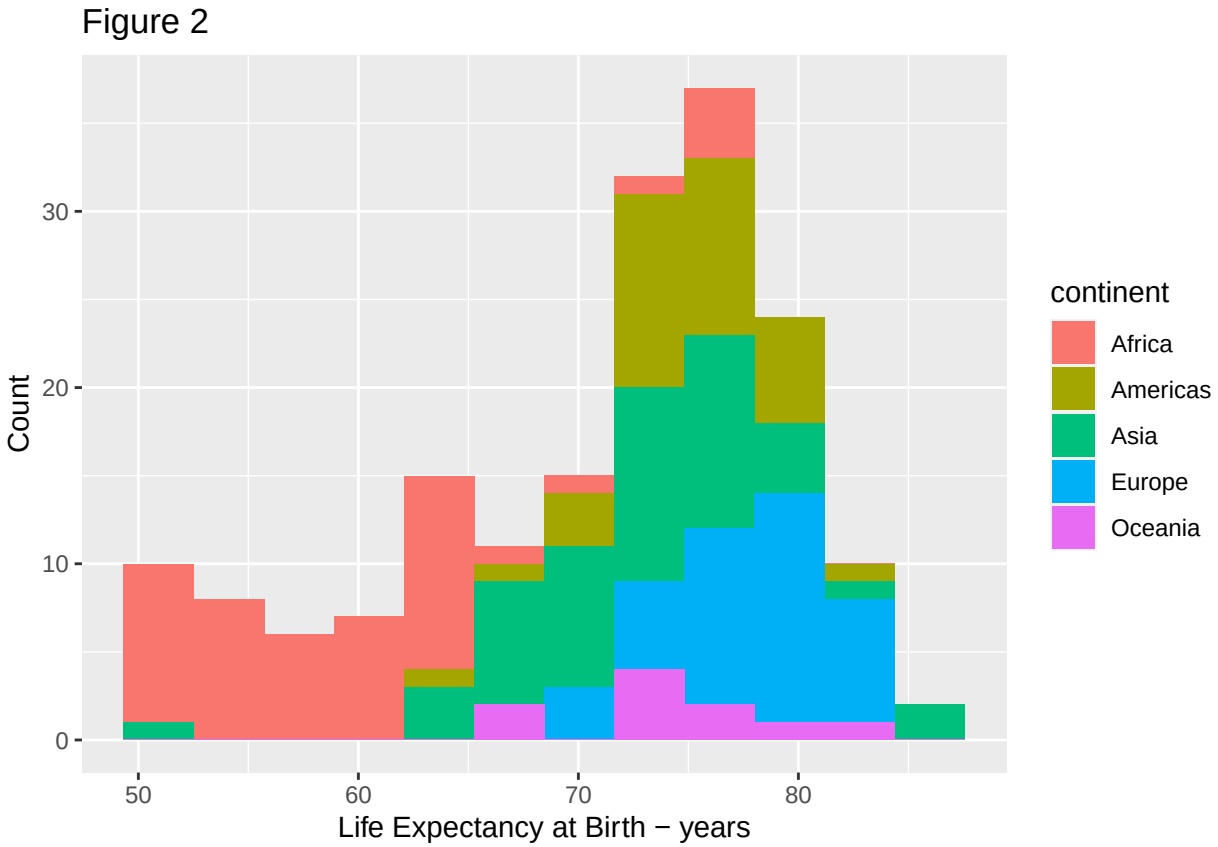
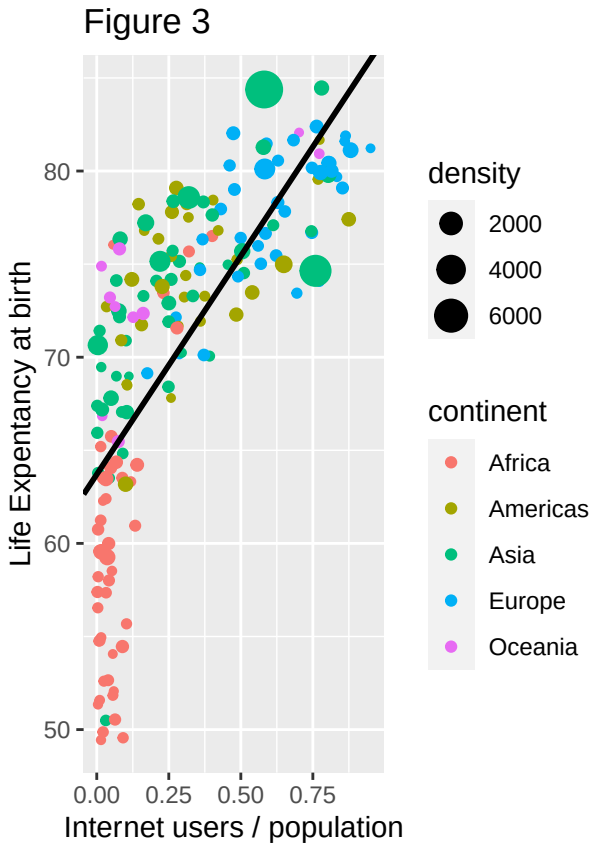


Figure 2: Africa takes most of the lower end of our life expectancy histogram, while the Americas, Asia, Europe, and Oceania carry similar ranges, with Asia as the lower of the bunch and Europe as the upper of the four. Life expectancy seems to have an upper skew across all countries, with the highest frequency occurring around the 75-76 range.



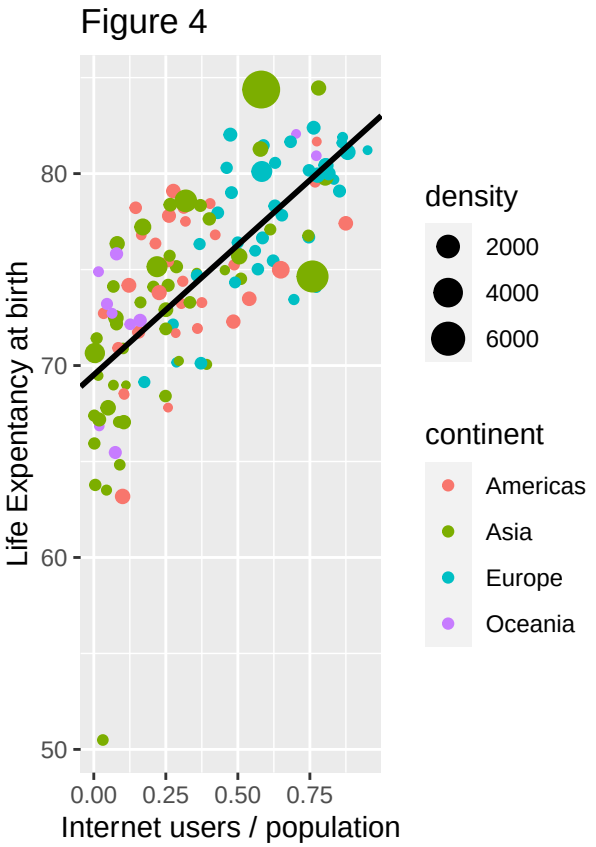
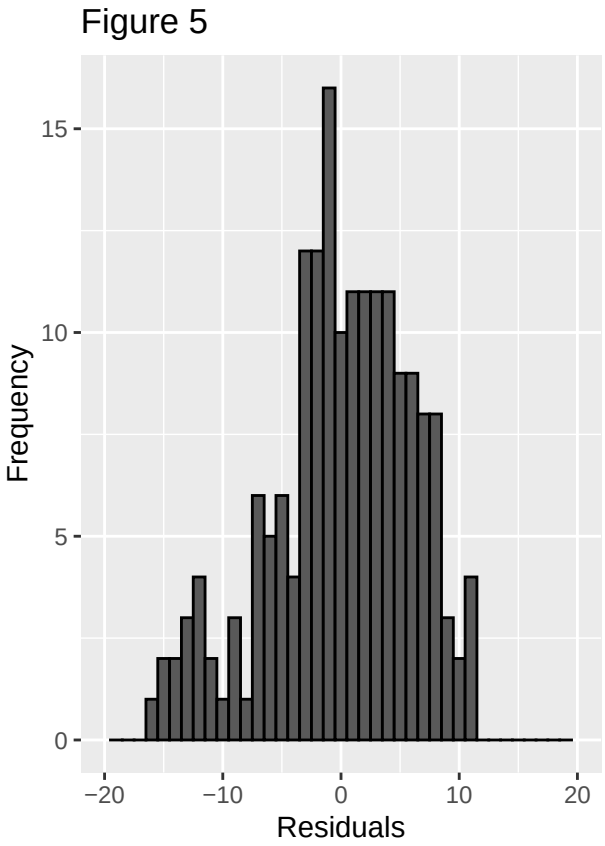


Figure 3: The proportion of internet users in the population of a given country correlates positively with life expectancy at birth. As indicated by the color —grouped by continent — these trends are not evenly dispersed across the nations of the world. Africa is clustered largely around the bottom left, with lower internet usership and life expectancy, while Europe is largely clustered in the top right, with higher internet usership and life expectancy. Oceania, Asia and the Americas are distributed across the plot.

Figure 4: After excluding Africa, a more linear correlation can be observed between life expectancy at birth and internet usage. There are significantly fewer outliers present than there were in the graph including Africa.



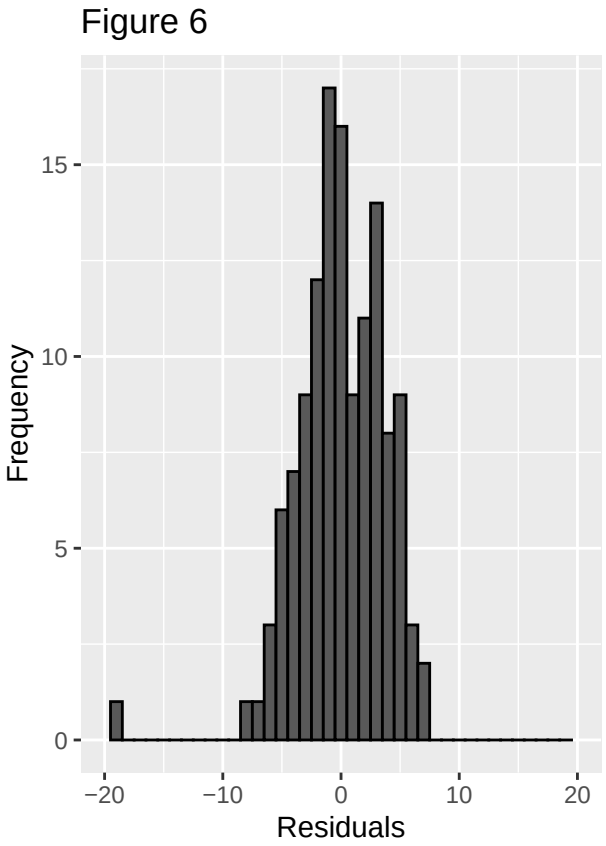


Figure 5: Our linear model tells us that for each 10% increase in the proportion of internet users, we can expect a 2.346 year increase in life expectancy. Our y-intercept within this model represents the life expectancy we should expect from a country with 0% internet usage, which here is estimated to be 63.71 years. R^2 within this model is 0.5291, which shows a minor correlation.

Our linear model including African countries is as follows:

$y_{lifeexpectancyatbirth} = B_0 + B_1 * x_{internetusership}$

$B_0 = 63.71$ $B_1 = 23.46$

Figure 6: Our linear model when excluding Africa tells us that for each 10% increase in the proportion of internet users, we can expect a 1.353 year increase in life expectancy. Our y-intercept within this model represents the life expectancy we should expect from a non-African country with 0% internet usage, which here is estimated to be 69.52 years. R^2 within this model is 0.4999, which shows a minor correlation surprisingly lower than our global model.

Our linear model excluding African countries is as follows:

$y_{lifeexpectancyatbirthwithoutAfrica} = B_0 + B_1 * x_{internetusership}$

$B_0 = 69.52$ $B_1 = 13.53$

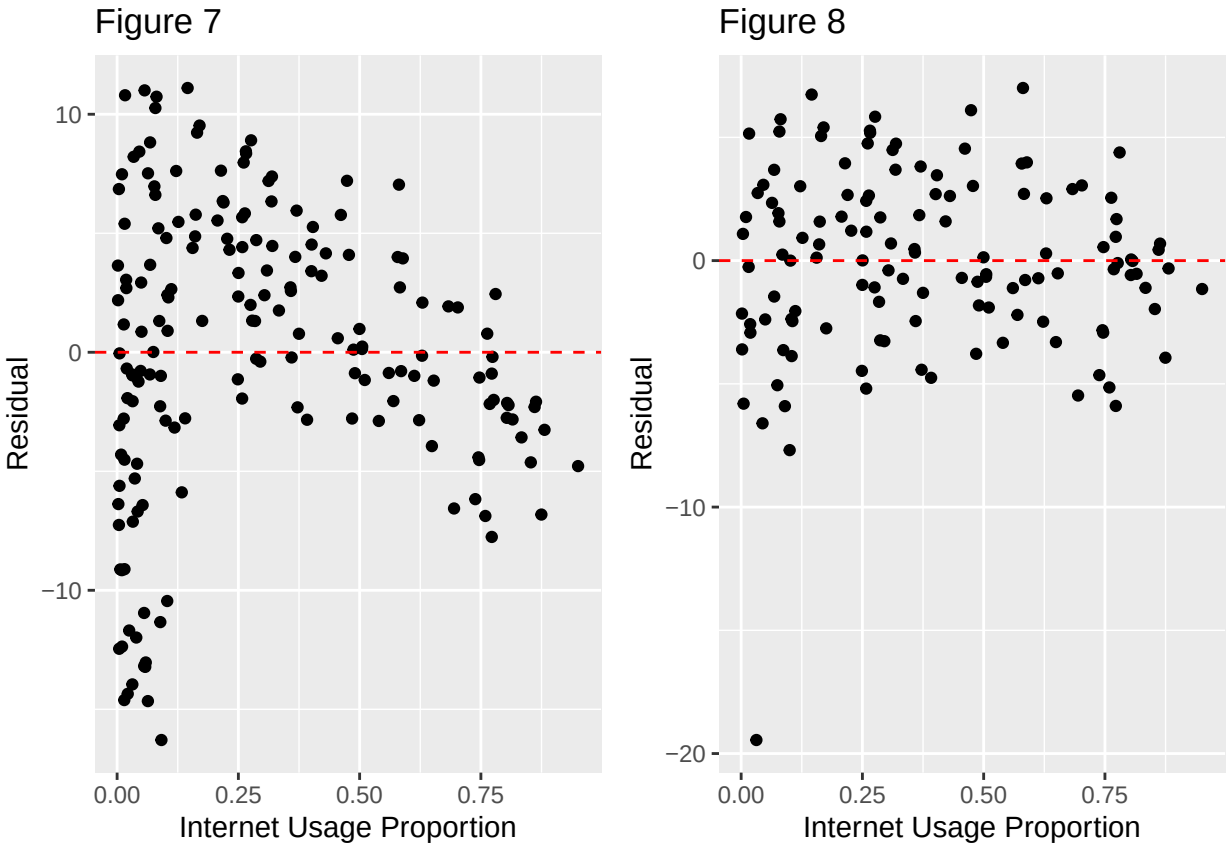


Figure 7: We can see that our linear model, in trying to accommodate both the trend among African countries and the trend among non-African countries, is divided (not as clearly unimodal) and shows an abnormal distribution of residuals. Since we are assuming normality for our simple linear model, we would expect this histogram to reflect a normal distribution, which it does not.

Figure 8: In excluding Africa, we see that our new linear model shows a more normal distribution of residuals, and so our linear model can more accurately describe the trend of our distribution.

Figure 9

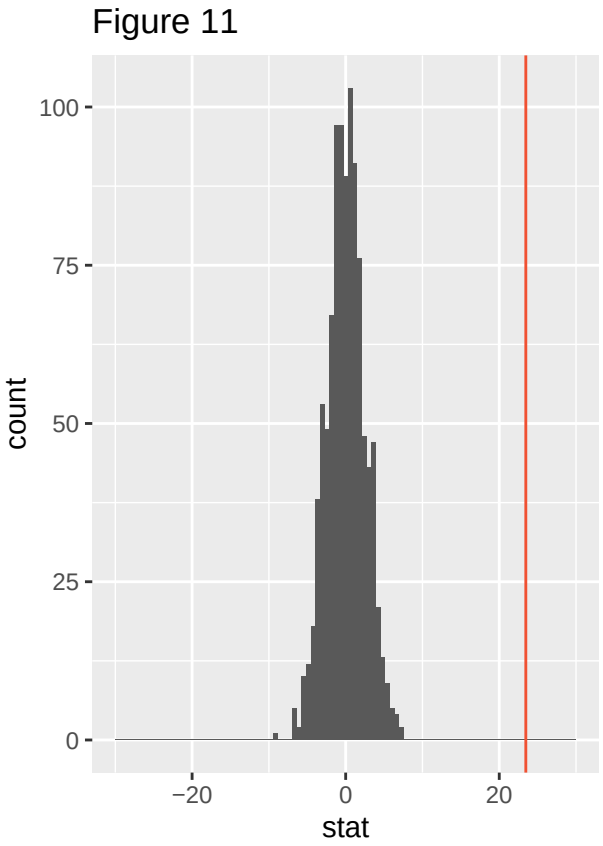
```
##
## Call:
## lm(formula = life_exp_at_birth ~ internet_usage_proportion, data = cia)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2903  -2.8736   0.1377   4.4186  11.1027
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 63.7069     0.6732   94.63   <2e-16 ***
## internet_usage_proportion  23.4594     1.6731   14.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.066 on 175 degrees of freedom
## Multiple R-squared:  0.5291, Adjusted R-squared:  0.5264
## F-statistic: 196.6 on 1 and 175 DF,  p-value: < 2.2e-16
```

Figure 9: This is a plot of residuals of the linear model with Africa included. If this was a good set of data to apply a linear model to, we would expect the points to be distributed normally around the 0 line. What we see in this scatterplot is clear heteroscedasticity and bias, indicating that this is not a good fit for a linear model although it yields a higher R^2 value.

Figure 10

```
##
## Call:
## lm(formula = life_exp_at_birth ~ internet_usage_proportion, data = noafrica)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.4508  -2.3693   0.0082   2.6411   7.0030
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 69.5157     0.5602  124.08   <2e-16 ***
## internet_usage_proportion  13.5291     1.2007   11.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.678 on 127 degrees of freedom
## Multiple R-squared:  0.4999, Adjusted R-squared:  0.496
## F-statistic:   127 on 1 and 127 DF,  p-value: < 2.2e-16
```

Figure 10: When we exclude Africa, we can see that our residuals lie within a consistent constrained range around our linear model. There is no obvious bias, and only a minimal noticeable amount of heteroscedasticity. This shows that our linear model is at the very least following a broader trend among our data points and that our slope is accurate.



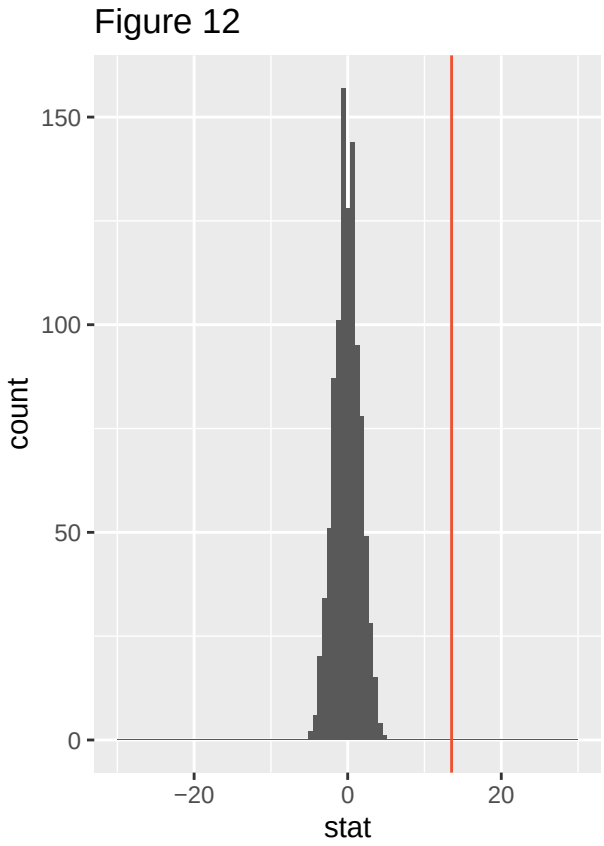


Figure 11: This is a histogram of slopes obtained from permuting the country data 1000 times and applying a linear model to each permutation. Of the simulated slopes, none approach the actual slope of the plot of the nations of the world. Accordingly, we are confident that we may reject the null hypothesis ($B_1 = 0$).

Figure 12: The exclusion of Africa does not change the fact that no simulated slopes approach the true slope of all non-African nations. Therefore, we can also reject the null hypothesis in the Africa excluded case.

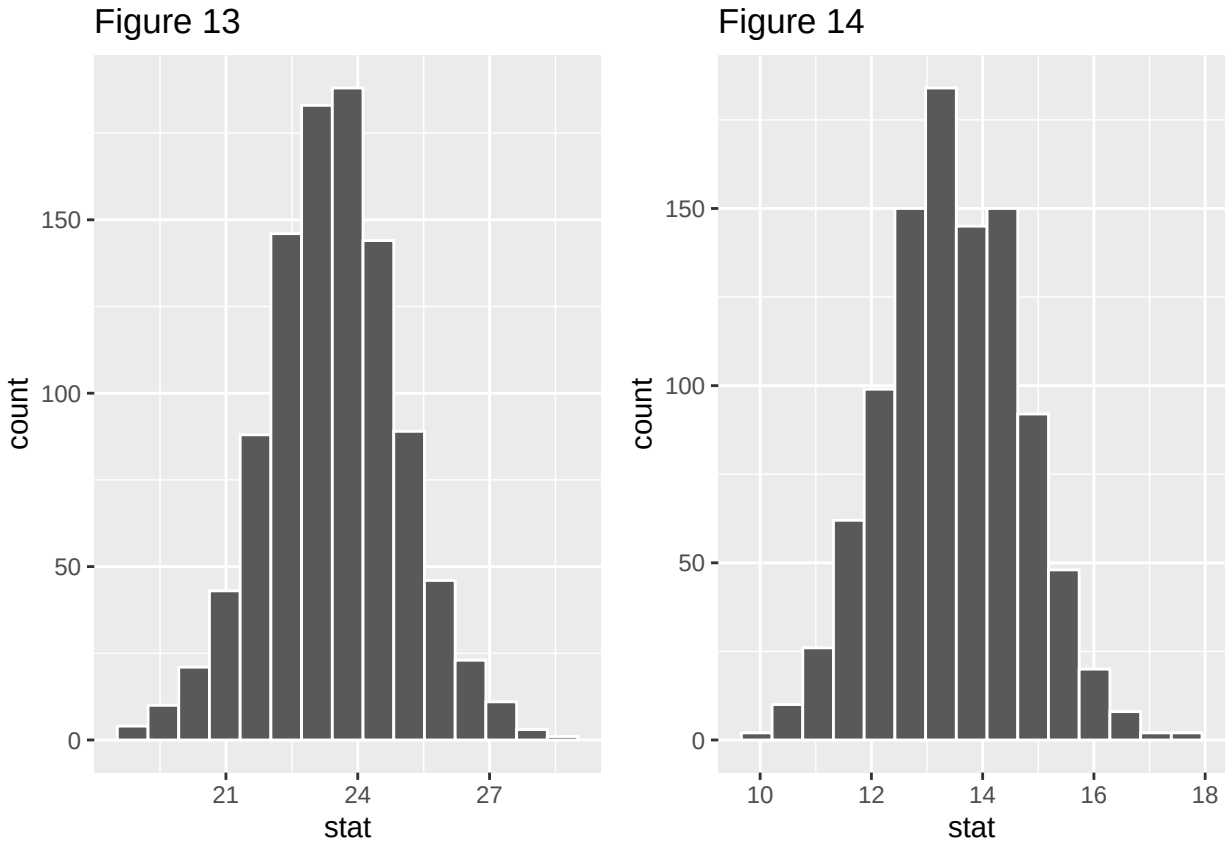


Figure 13: This histogram of slopes is obtained from bootstrapping our data and fitting slopes to the bootstrapped samples. It is centered around ~23, our observed slope with Africa included.

Figure 14: This histogram of slopes is obtained from bootstrapping our data and fitting slopes to the bootstrapped samples. It is centered around ~13, our observed slope with Africa excluded.

Figure 15

```
##     2.5%    97.5%
## 20.39227 26.50567
```

Figure 15: We are 95% confident that the slope of life expectancy at birth vs internet usership is between 20.392 and 26.506.

Figure 16

```
##     2.5%    97.5%
## 11.10236 15.85847
```

Figure 16: We are 95% confident that the slope of life expectancy at birth vs internet usership in nations excluding those of Africa is between 11.102 and 15.859.

> *Note: In all of our scatterplots, the dots representing countries have size indicating their population density. We used this technique to explore multiple aspects of each country's data, and it doesn't have a direct bearing on our analysis.*

Conclusions

We observed a positive correlation between internet usership in a given country, and life expectancy at birth. While this relationship varies across the nations and continents of the world – with much of Africa serving as a noted exception – we are 95% confident that the true slope of this data set is between 20.392 and 26.506. This is sufficient evidence to reject the null hypothesis. Our analysis of the residuals, however, tells us that this relationship might be trying to accommodate a non-linear relationship, or possibly two separate linear relationships. Exploring the second possibility, at the same confidence interval this window is lowered to between 11.103 and 15.858. This is still sufficient to reject the null hypothesis, and our residual analysis shows a significantly more normal distribution than before. This points to the linear model fitting the non-African relationship more comfortably than the global relationship between life expectancy at birth and proportion of internet users.

Possible shortcomings include the employment of the linear model, which was not perfectly suited to the data, at least not when applied to all nations of the world. A linear model may, however, be applied to non-African nations, with our residual analysis explaining their applicability. Simply excluding Africa from our analysis yielded fruitful results, but our analysis falls short of telling us what is unique about Africa in this respect. Further research into this topic would benefit from more advanced models, more research from the social sciences explaining external influences, or access to other variables like level of economic development, geography, or urbanization. Understanding the differences belying the African and non-African divide could point us to whether life expectancy or internet usage is the true explanatory variable.

References

1. Lee, C.-W., & Kim, M.-S. (2018). The Relationship between Internet Environment and Life Expectancy in Asia. Review of Integrative Business and Economics Research, 8(2), 70–80.
2. Alzaid, A., Alsulami, M., Komal, K., & Al-Maraghi, A. (n.d.). 24th IBIMA Conference. In Examining the Relationship between the Internet and Life Expectancy (pp. 1–10). Milan.