```
cia <- readr::read_csv("https://raw.githubusercontent.com/nichollsharrisonj/Stats-Project-1/master/cia_
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   country = col_character(),
##   area = col_double(),
##   birth_rate = col_double(),
##   death_rate = col_double(),
##   infant_mortality_rate = col_double(),
##   internet_users = col_double(),
##   life_exp_at_birth = col_double(),
##   maternal_mortality_rate = col_double(),
##   net_migration_rate = col_double(),
##   population = col_double(),
##   population_growth_rate = col_double()
## )
```

```
cia <- cia %>% mutate(
  continent = countrycode(country,origin="country.name",destination="continent"),
  density = (population/area),
  internet_usage_proportion = internet_users/population
)
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : Some val
```

```
cia <- na.omit(cia) #OMIT ALL ROWS WITH ANY NA VALUES
```

Statistics Project – Math 141

Dashiell Ward, Gavin Rimmer, Bailee Brunsmann, Harrison Nicholls

CIA Factbook

Question:

In a given nation, does the proportion of the population using the internet correlate to life expectancy at birth?

Hypothesis:

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

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

Technical Conditions

Linearity. The scatterplot of the explanatory and response must be nearly linear.

Independent Observations. The samples must be independent.

Normally Distributed Residuals. The errors must show a nearly normal distribution.

Constant or equal variability. The error must exhibit homoscedasticity.

Dataset: What is it?

- Details on countries

Where does it come from?

- CIA Factbook

What is the description of each variable?

- See 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.

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

Further Research:

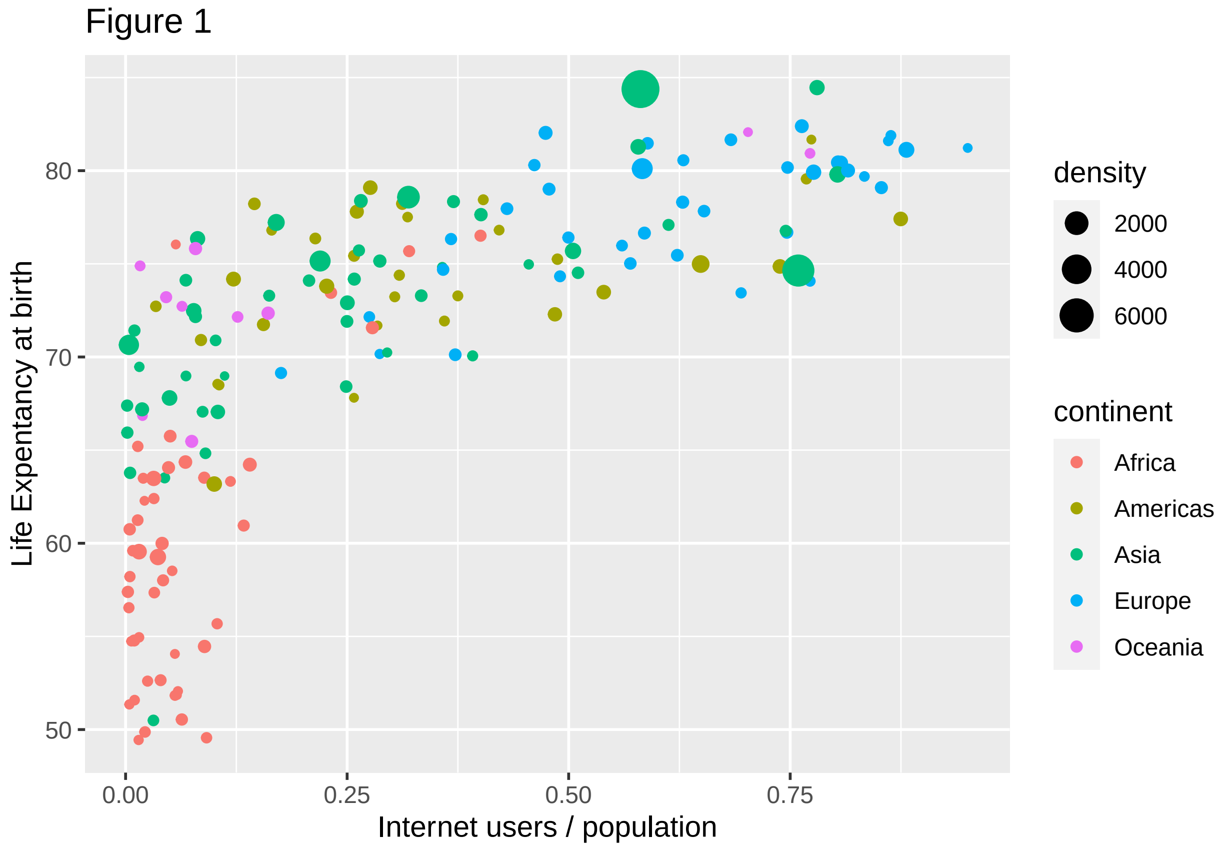
Lee, Cheng-Wen, The Relationship between Internet Environment and Life Expectancy in Asia

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.

Exploratory plots:



```
life_exp <- cia$life_exp_at_birth
s <- sd(life_exp)
se <- s/sqrt(length(life_exp))
se
```

```
## [1] 0.6625533
```

```
t <- 23.459 / se
t
```

```
## [1] 35.40696
```

```
df = length(life_exp) - 2
df
```

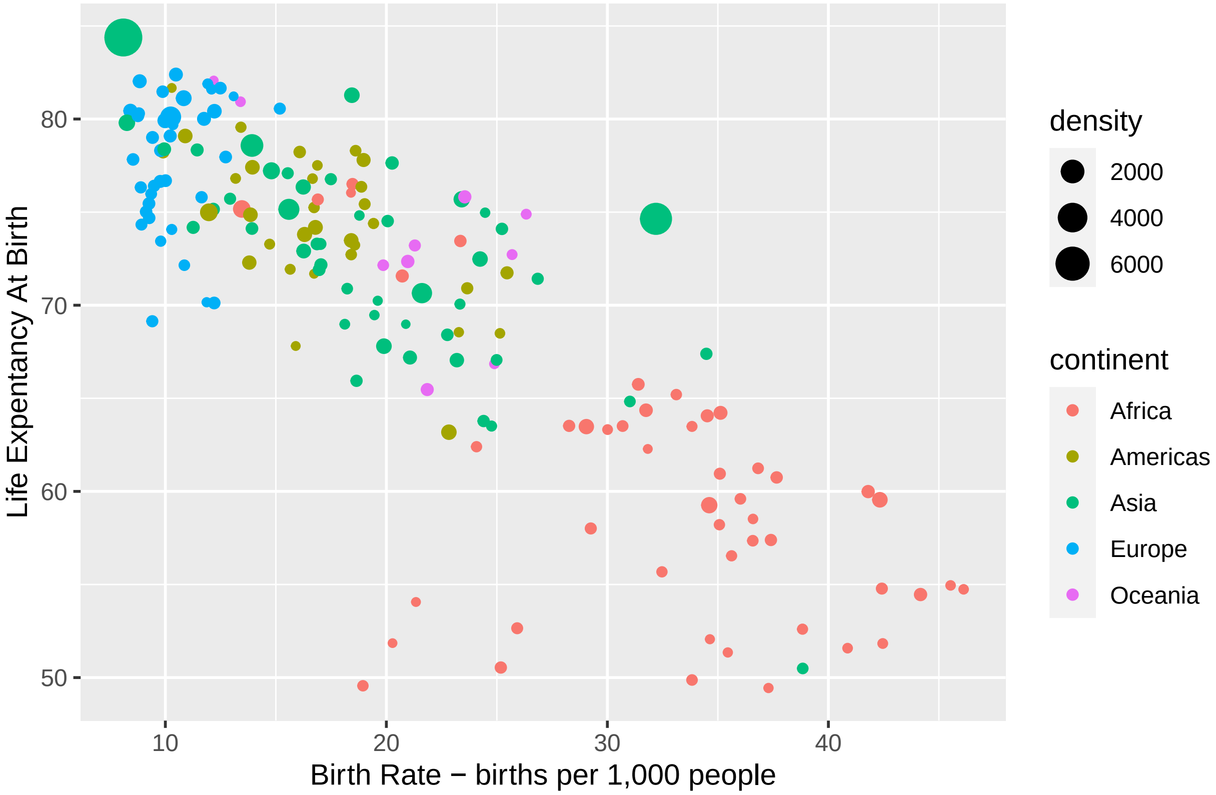
```
## [1] 175
```

```
2*qt(t,df)
```

```
## Warning in qt(t, df): NaNs produced
```

```
## [1] NaN
```

Figure 2



Life Expectancy At Birth
80
70
60
50
10
20
30
40
Birth Rate − births per 1,000 people
density
● 2000
● 4000
● 6000
continent
● Africa
● Americas
● Asia
● Europe
● Oceania

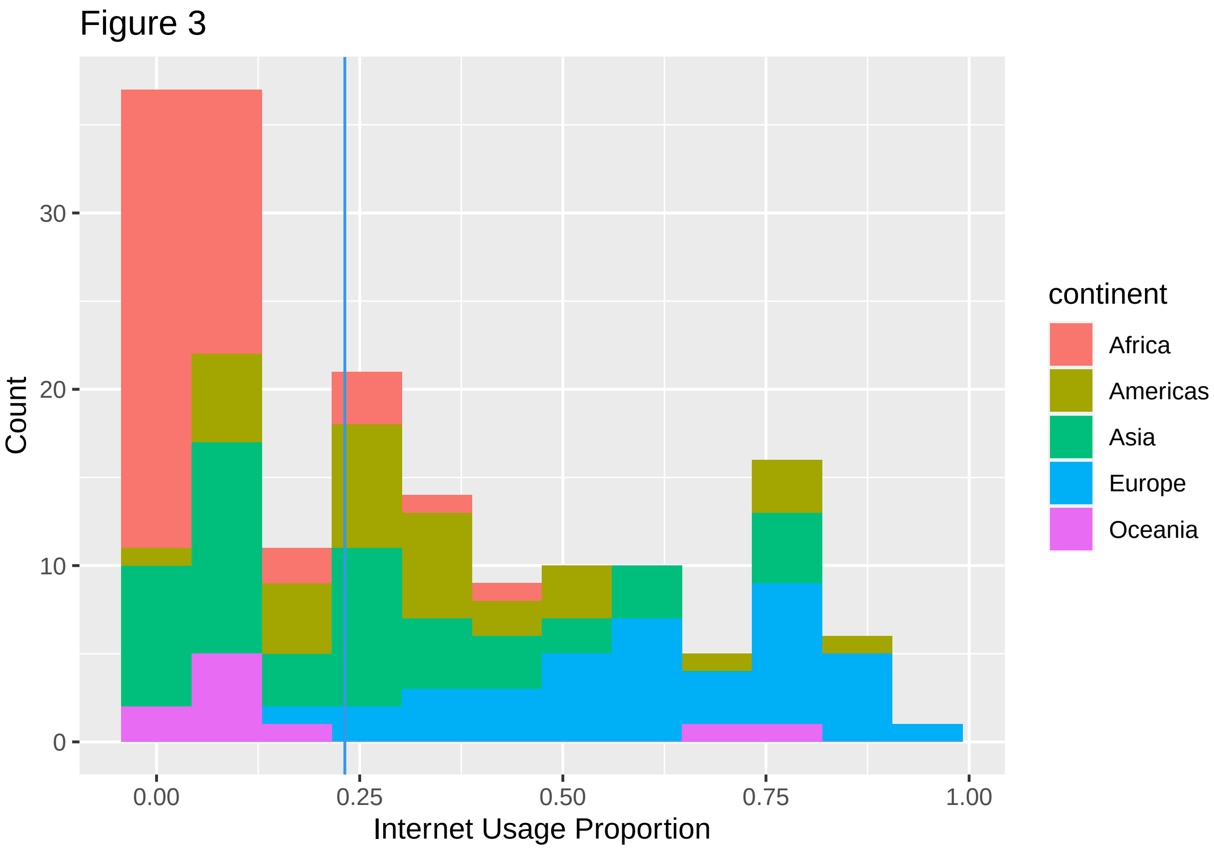


Figure 3
Count
30
20
10
0
0.00
0.25
0.50
0.75
1.00
Internet Usage Proportion
continent
Africa
Americas
Asia
Europe
Oceania

Figure 4

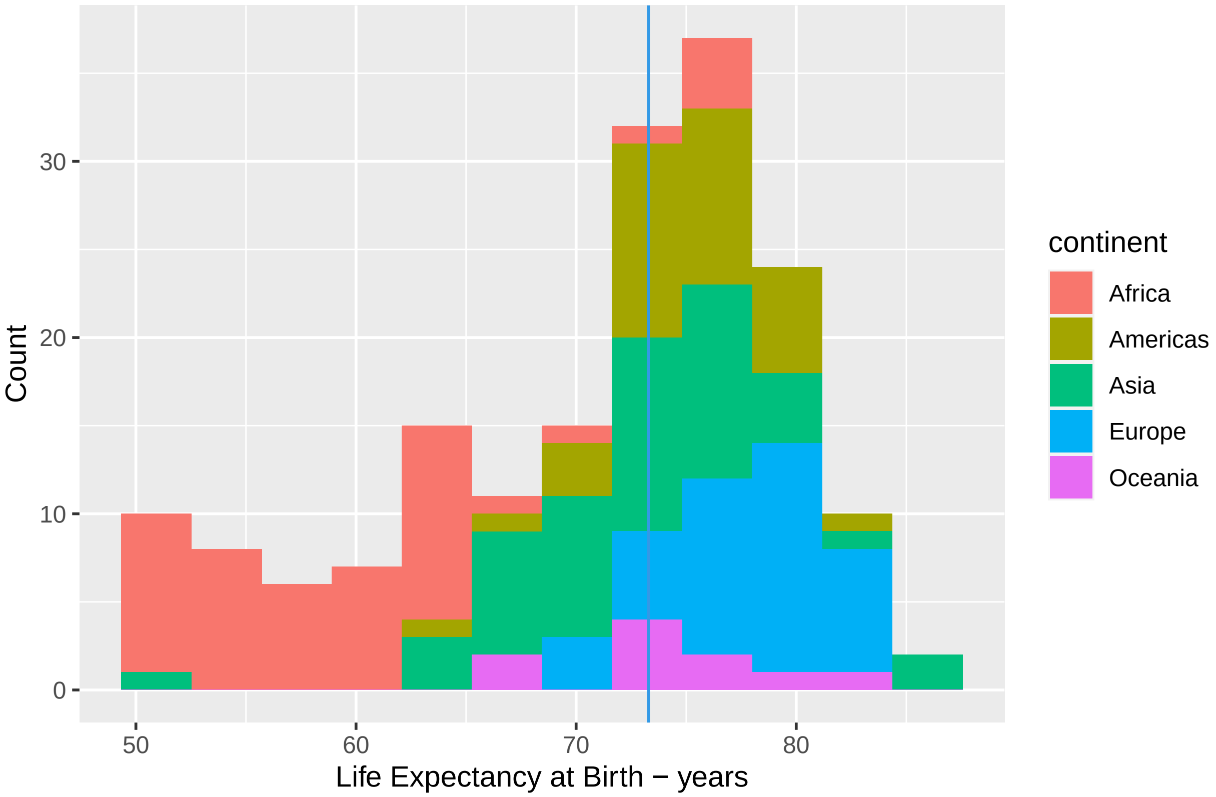

Count
30
20
10
0
50
60
70
80
Life Expectancy at Birth − years
continent
Africa
Americas
Asia
Europe
Oceania

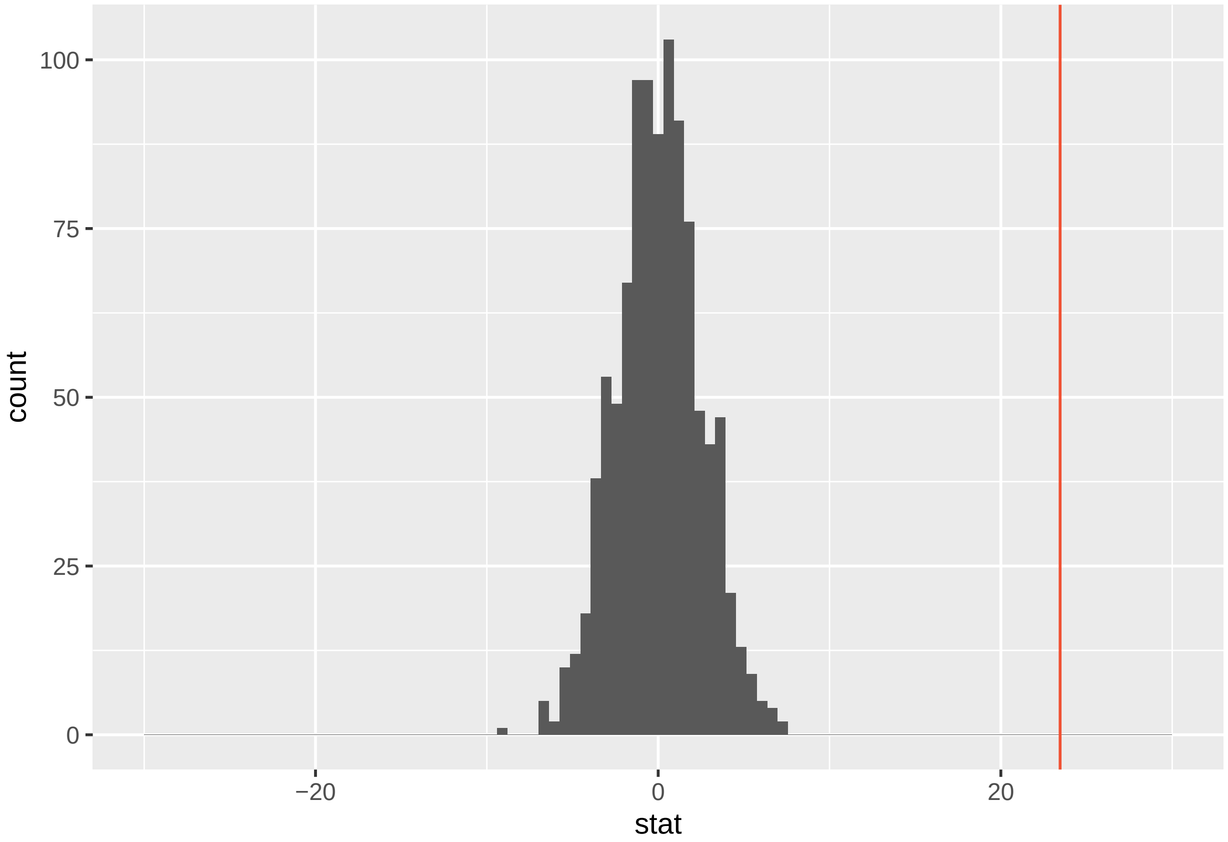
Figure 6 - Linear model with Africa included

```
lm_mod <- lm(life_exp_at_birth ~ internet_usage_proportion, data = cia)

summary(lm_mod)
```

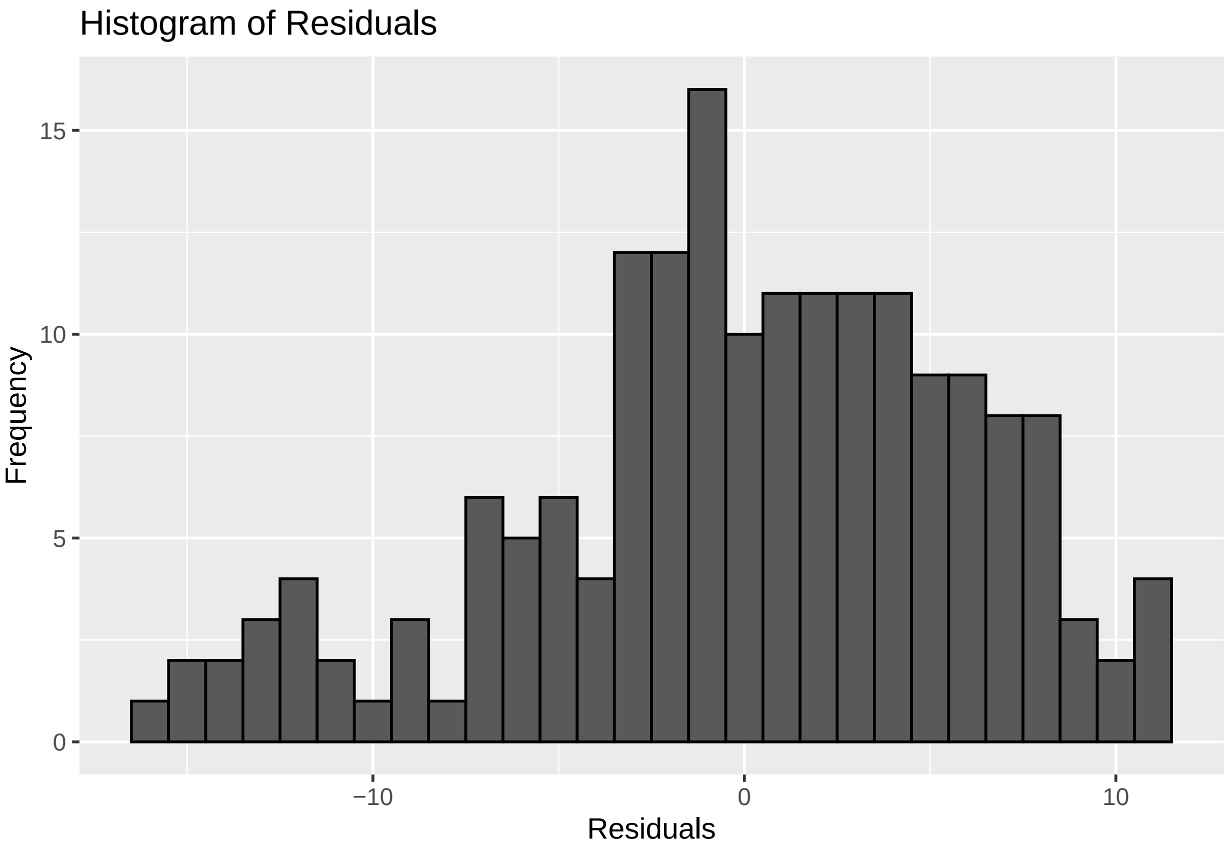
```
##
## Call:
## lm(formula = life_exp_at_birth ~ internet_usage_proportion, data = cia)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2903  -2.8736   0.1377   4.4186  11.1027
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                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
```

```
obs_slope <- lm_mod$coefficients[2]
```

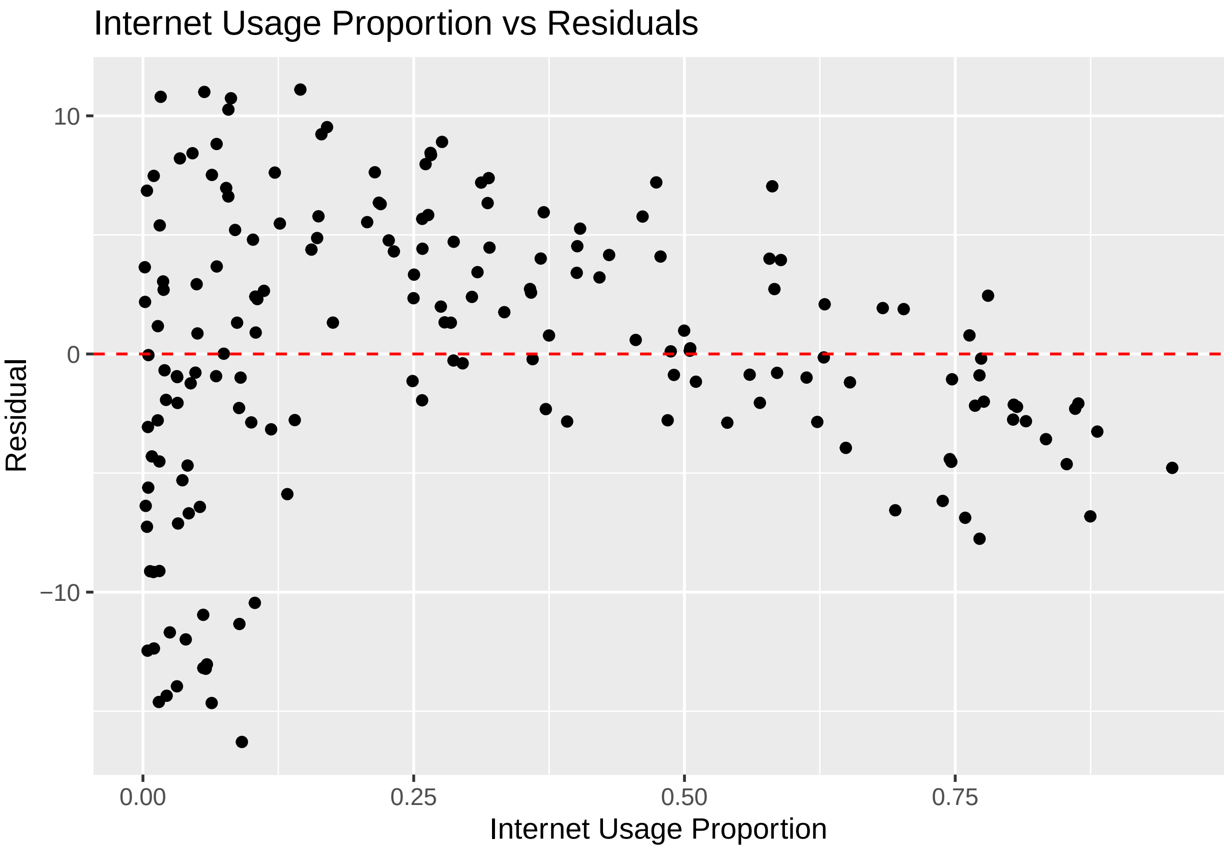
```
set.seed(1009)
perm_slope <- cia %>%
   specify(life_exp_at_birth ~ internet_usage_proportion) %>%
   hypothesize(null = "independence") %>%
   generate(reps = 1000, type = "permute") %>%
   calculate(stat = "slope")

ggplot(data=perm_slope, aes(x=stat)) +
   geom_histogram(bins=100) +
   geom_vline(xintercept = obs_slope,
              color = "#F05133") +
  xlim(-30,30)
```



```
lm_res <- resid(lm_mod)
lm_res <- data.frame(resid = lm_res)
lm_res <- lm_res %>% mutate(internet_usage_proportion = cia$internet_usage_proportion)
histogram <- ggplot(data=lm_res, aes(x=resid))
histogram + geom_histogram(binwidth=1, color="black") +
  xlab("Residuals") +  ylab("Frequency") + ggtitle("Histogram of Residuals")
```

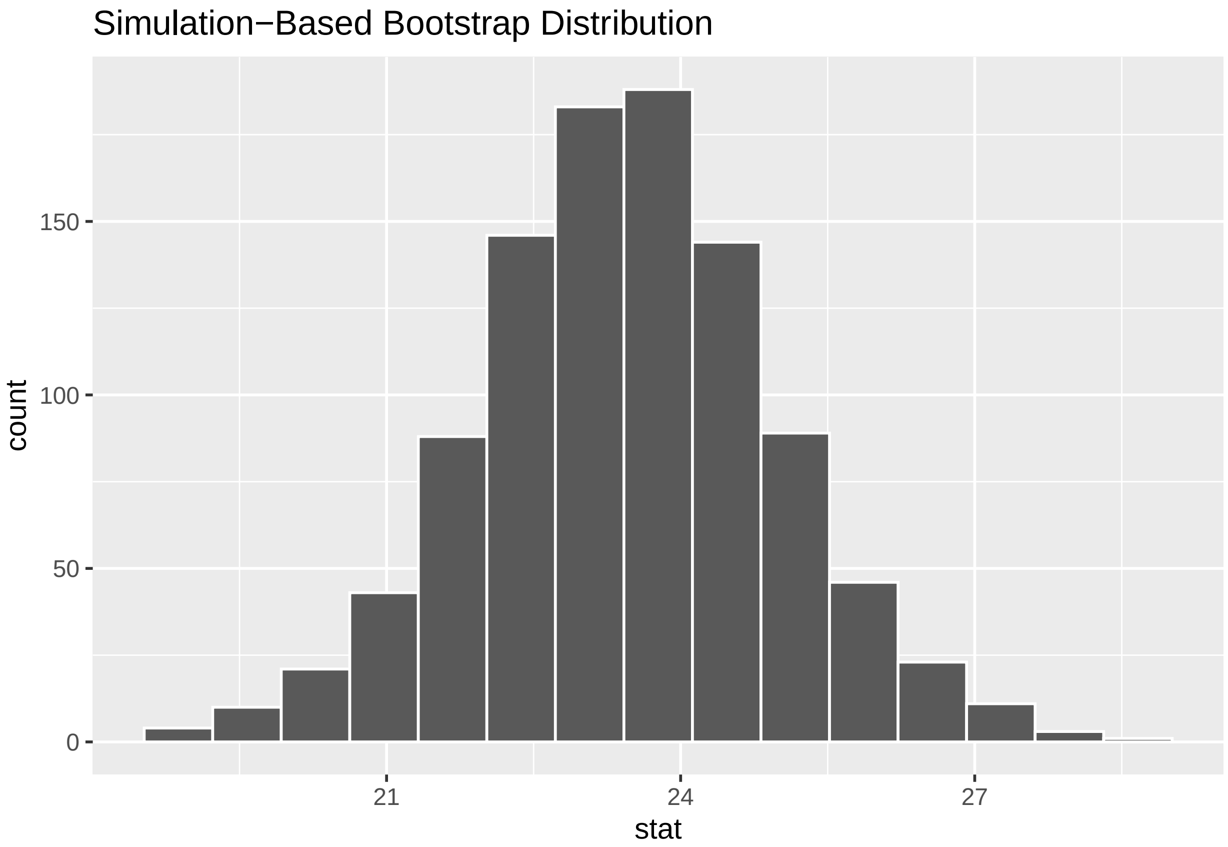


```
scatter <- ggplot(data=lm_res, aes(x = internet_usage_proportion, y =resid))
scatter + geom_point(size = 1.5) +
  xlab("Internet Usage Proportion") +  ylab("Residual")  +
  geom_hline(yintercept=0, linetype='dashed', col = 'red') +
  ggtitle("Internet Usage Proportion vs Residuals")
```



The observed slope is nowhere near our slopes obtained by permuting the data. we are certain that we can reject the null hypothesis.

```
set.seed(34209)
perm_ci <- cia %>%
  specify(life_exp_at_birth ~ internet_usage_proportion) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "slope")
visualize(perm_ci)
```



```r
alpha <- .05

#lower percentile cutoff
p_lower <- .025

#upper percentile cutoff
p_upper <- 1-(.025)

# Create a confidence interval of stat using quantiles
quantile(perm_ci$stat,c(p_lower,p_upper))
```

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

Our 95% confidence interval for slope of the data with Africa included is (20.392,26.506). The null slope of 0 is not included within this confidence interval.

```
noafrica <- cia %>%
  filter(continent != "Africa")

ggplot (data=noafrica, aes(x=internet_usage_proportion, y=life_exp_at_birth)) +
  geom_point(aes(color = continent,size=density)) +
  xlab('Internet users / population') +
  ylab('Life Expentancy at birth') +
  ggtitle('Figure 7 - Africa Excluded')
```

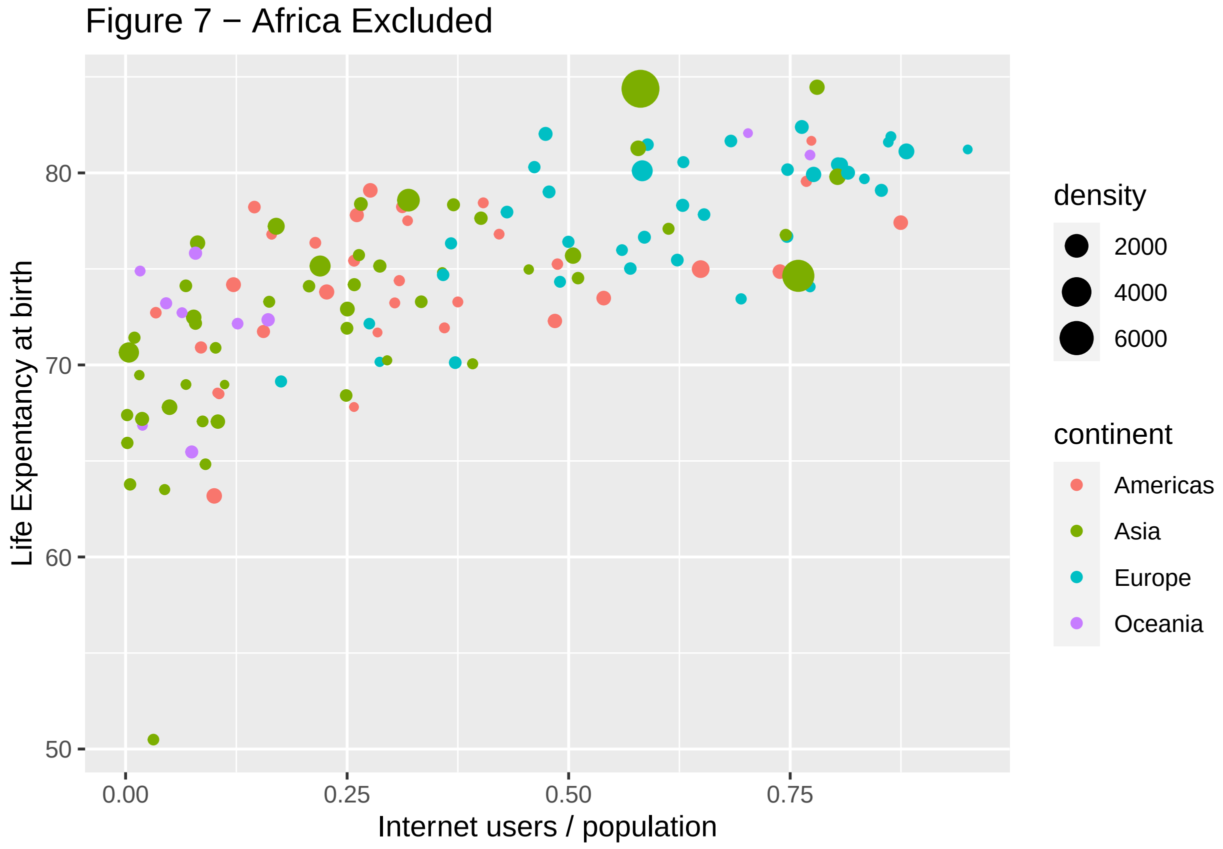


Figure 7 − Africa Excluded

Figure 8 - Linear model with Africa Excluded

```
lm_mod_noafrica <- lm(life_exp_at_birth ~ internet_usage_proportion, data=noafrica)
summary(lm_mod_noafrica)
```

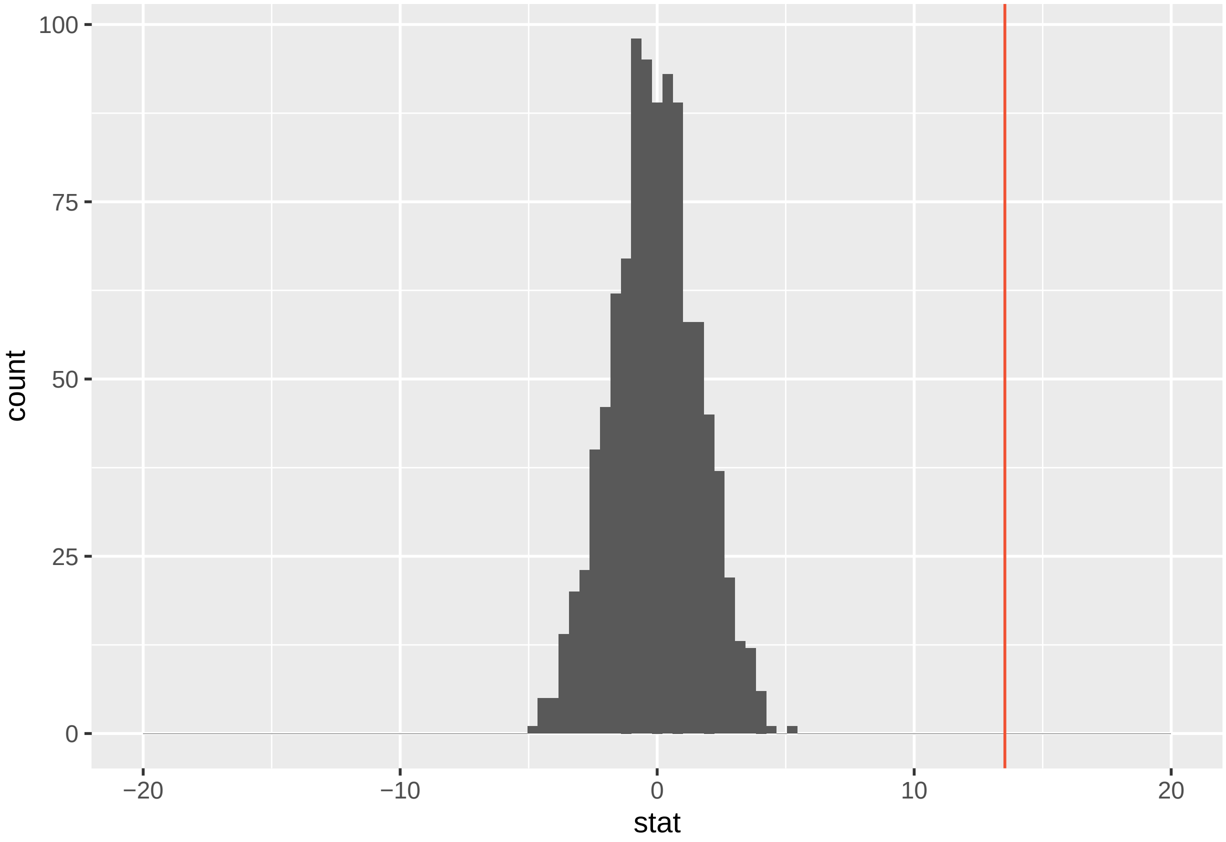
```
##
## Call:
## lm(formula = life_exp_at_birth ~ internet_usage_proportion, data = noafrica)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.4508  -2.3693   0.0082   2.6411   7.0030
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                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
```

```
obs_slope_noafrica <- lm_mod_noafrica$coefficients[2]

obs_slope_noafrica
```

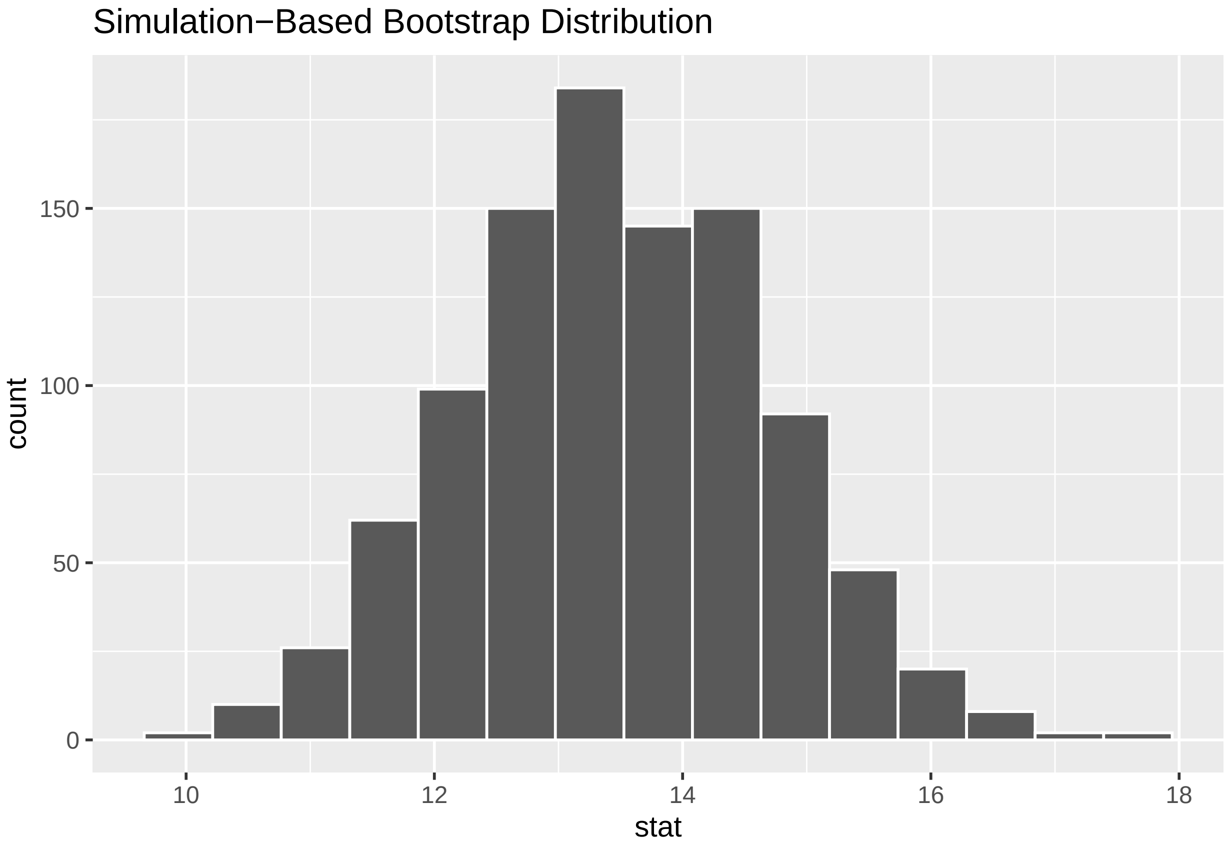
```
## internet_usage_proportion
##                  13.52906
```

Inference with no Africa

```
set.seed(1010)
perm_slope_noafrica <- noafrica %>%
   specify(life_exp_at_birth ~ internet_usage_proportion) %>%
   hypothesize(null = "independence") %>%
   generate(reps = 1000, type = "permute") %>%
   calculate(stat = "slope")

ggplot(data=perm_slope_noafrica, aes(x=stat)) +
   geom_histogram(bins=100) +
   geom_vline(xintercept = obs_slope_noafrica,
              color = "#F05133") +
  xlim(-20,20)
```



```r
set.seed(34210)
perm_ci_noafrica <- noafrica %>%
  specify(life_exp_at_birth ~ internet_usage_proportion) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "slope")
visualize(perm_ci_noafrica)
```

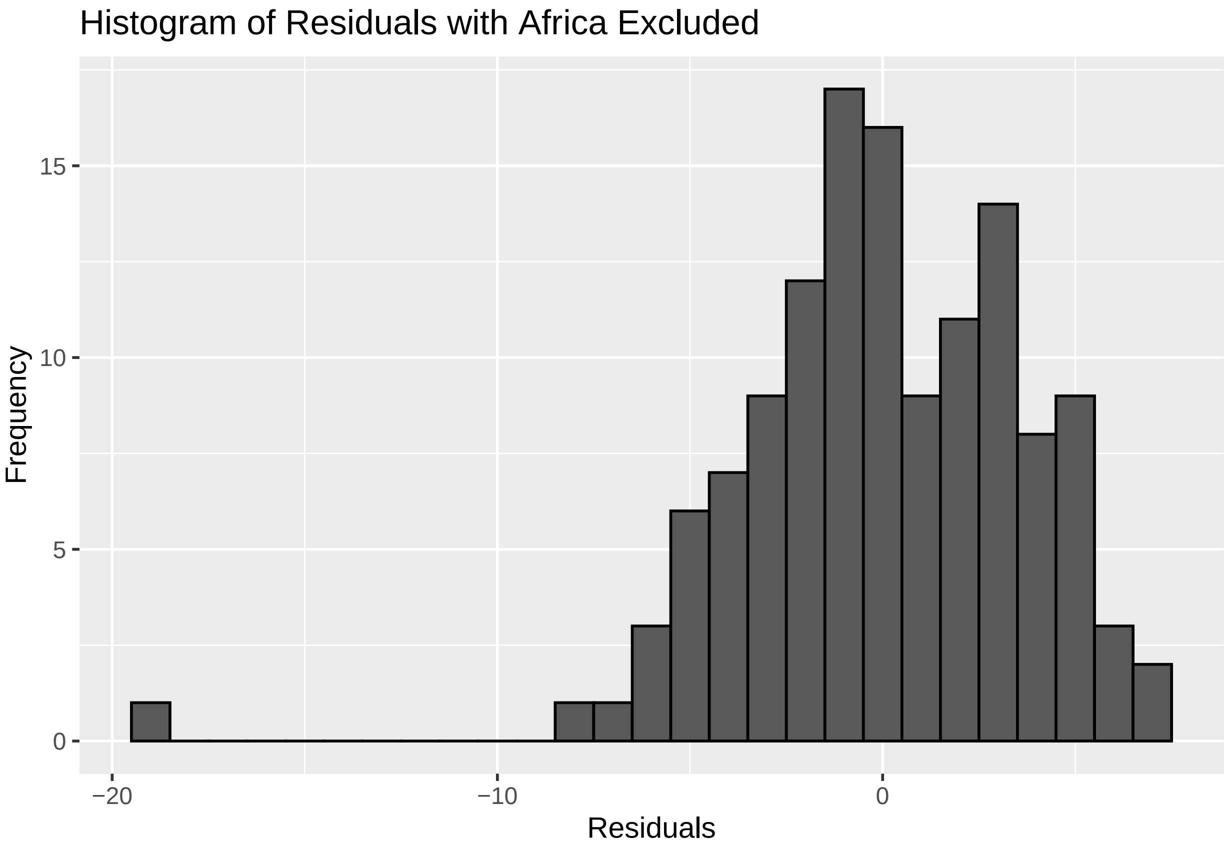


```r
alpha <- .05

#lower percentile cutoff
p_lower <- .025

#upper percentile cutoff
p_upper <- 1-(.025)

# Create a confidence interval of stat using quantiles
quantile(perm_ci_noafrica$stat,c(p_lower,p_upper))
```

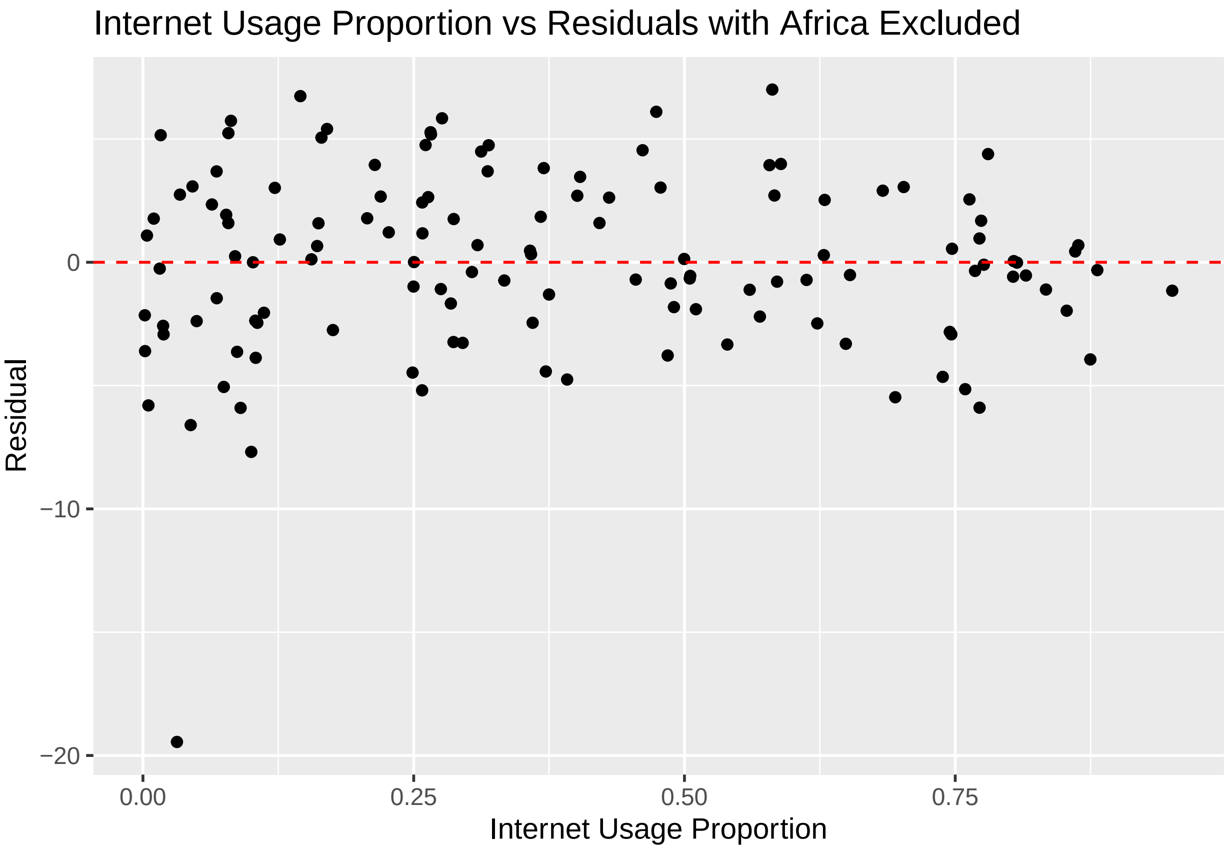
```
##     2.5%    97.5%
## 11.10236 15.85847
```

~~~
```r
lm_res_noafrica <- resid(lm_mod_noafrica)
lm_res_noafrica <- data.frame(resid = lm_res_noafrica)
lm_res_noafrica <- lm_res_noafrica %>% mutate(internet_usage_proportion = noafrica$internet_usage_propo
histogram <- ggplot(data=lm_res_noafrica, aes(x=resid))
histogram + geom_histogram(binwidth=1, color="black") +
~~~

```
  xlab("Residuals") +  ylab("Frequency") + ggtitle("Histogram of Residuals with Africa Excluded")
```



```
scatter <- ggplot(data=lm_res_noafrica, aes(x = internet_usage_proportion, y =resid))
scatter + geom_point(size = 1.5) +
  xlab("Internet Usage Proportion") +  ylab("Residual")  +
  geom_hline(yintercept=0, linetype='dashed', col = 'red') +
  ggtitle("Internet Usage Proportion vs Residuals with Africa Excluded")
```


Internet Usage Proportion vs Residuals with Africa Excluded
Residual
0
−10
−20
0.00
0.25
0.50
0.75
Internet Usage Proportion

Statistical Methods:

In performing an observational study, we are limited in the viability of performing randomization tests on our data. We can find correlations in our plots using Rˆ2 and find the differences in means between different categorical groups, in our case Continents seems apt for a geopolitical analysis of the data. However, given that our sample population is itself the true population of all countries, there are no methods by which we can fabricate more samples via bootstrapping.

Figure Five: Here we have found the mean birth rate of each continent, and each mean's difference from the global mean. From this we can glean that the Americas, Asia, and Oceania all lie very close to the global mean, while Africa and Europe both fall in the upper and lower end respectively.

We 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. We therefore theorized that internet usage only begins to correlate linearly with life expectancy at some threshold value (~65 years), which hardly any African countries meet. Based on this, we decided to see how the data would look with Africa excluded. Figures 7 and 8 reflect the results of this exploration.

Figure Six, Figure Seven, and Figure Eight: Here we have generated linear models based on Figure 1, and computed Rˆ2 of each linear model within the graphs. The greatest Rˆ2 value, and thereby the best-fit linear model, is calculated to be that of Figure Six, the graph which includes all countries, rather than Figure Seven or Figure Eight which exclude countries in the continent of Africa.

To further explore our findings, we plan to make plots of the residuals and include lines of best fit. The residuals will give us insight into the biases in our scatterplots, and allow us to determine how appropriate linear models are for comparing internet usage and life expectancy. Although excluding Africa yields a worse Rˆ2 value, it may be that the residuals form a distribution closer to normal, meaning that it is more suited to a linear model than the alternative.

Analysis of Data Visualizations:

Figure One: The proportion of internet users in the population of a given country correlates positively with life expectancy at birth. As indicated by the shading — indicating 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 Two: Birth rate appears to be inversely proportional to life expectancy at birth, meaning that individuals in countries with higher life expectancy tend to have fewer children. As indicated by the shading — indicating continent — these trends are not evenly dispersed across the nations of the world. Africa and Europe again find themselves at opposing ends of the plot, with higher birth rates, lower life expectancy, and lower birth rates, higher life expectancy respectively. Once more, Oceania, Asia, and the Americas lie dispersed in the middle.

Figure Three: Africa and Europe take the upper end of our death rate histogram, with Asia, Oceania, and the Americas occupying the lower end. Africa seems to have the largest range in death rate. The graph takes a minor lower skew, though it makes an asymmetrical peak in greatest frequency around 8 deaths per 1,000 population.

Figure Four: 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 7: In excluding Africa, we are provided with a new angle on Figure One. While the relationship is the same, bias appears lower. Despite this, the correlation without Africa turns out to be lower than when Africa is included.

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