Corruption Regression

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

```
library(haven)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#install.packages("stargazer")
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
#install.packages("car")
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.3.2
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
#install.packages("readxl")
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.3.2
```

```
getwd()
```

```
## [1] "C:/Users/chris/OneDrive/Documents/ECO 4421/Regression_Project"
```

```
"C:\\Users\\chris\\OneDrive\\Documents\\ECO 4421\\Regression_Project\\corruption_index_clean_data.xlsx"
```

```
## [1] "C:\\Users\\chris\\OneDrive\\Documents\\ECO 4421\\Regression_Project\\corruption_index_clean_data.xlsx"
```

```
corruption_data <- read_excel("corruption_index_clean_data.xlsx")
head(corruption_data)
```

country	corruption_index	annual_income	tourists_in_millions	cost_index
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
Denmark	12	68110	NA	119.9
Finland	12	53660	NA	108.0
New Zealand	12	45340	NA	117.2
Norway	15	84090	1.4	124.6
Singapore	15	64010	NA	75.0
Sweden	15	58890	NA	109.3

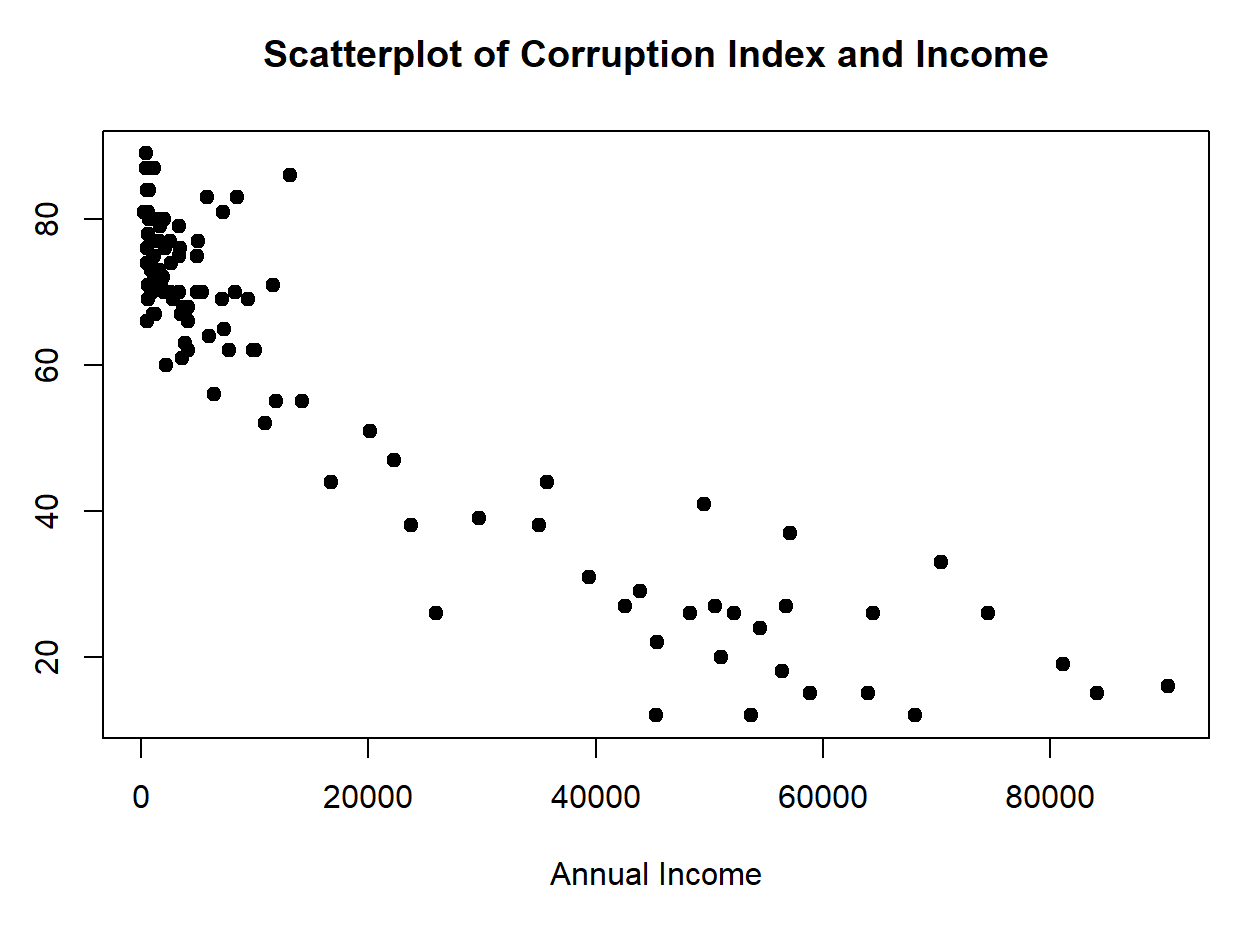
6 rows

```
glimpse(corruption_data)
```

```
## Rows: 110
## Columns: 5
## $ country              <chr> "Denmark", "Finland", "New Zealand", "Norway", "S…
## $ corruption_index     <dbl> 12, 12, 12, 15, 15, 15, 16, 18, 19, 20, 22, 24, 2…
## $ annual_income        <dbl> 68110, 53660, 45340, 84090, 64010, 58890, 90360, …
## $ tourists_in_millions <dbl> NA, NA, NA, 1.4, NA, NA, NA, 7.3, 0.5, 12.4, NA, …
## $ cost_index           <dbl> 119.9, 108.0, 117.2, 124.6, 75.0, 109.3, 142.4, 9…
```

Off the bat, from the scatterplot, we can see a clear cluster of countries that have a high corruption index tend to have citizens with income below ~10,000.

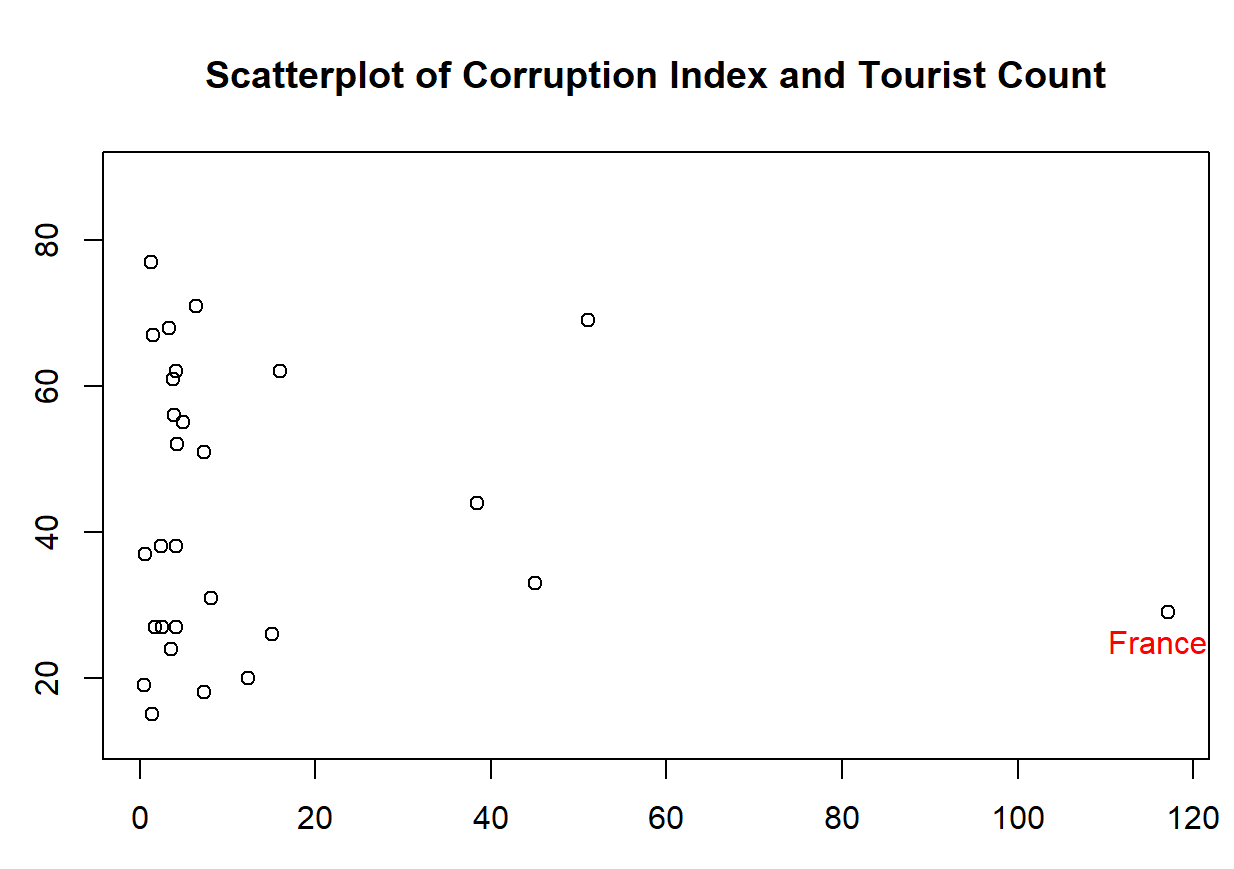
```
plot(corruption_data$annual_income, corruption_data$corruption_index, pch=19, xlab = "Annual Income", ylab = "Cor
ruption Index", main = "Scatterplot of Corruption Index and Income")
```



Scatterplot analyzing corruption index

and tourist count seems to reveal no discernable and clear relationship between the two, though it does reveal an outlier in tourist count with relatively low corruption index, France. Nonetheless, it should still be included in our final regression as a control variable to potentially account for other factors correlated with annual income and determinants of corruption index (eg. perceived safety, political stability, crime rates).

```
plot(corruption_index ~ tourists_in_millions,
     data = corruption_data,
     main = "Scatterplot of Corruption Index and Tourist Count",
     xlab = "Tourist Count",
     ylab = "Corruption Index")
text(116, 25, "France", col = "red")
```



Regressing only annual income on

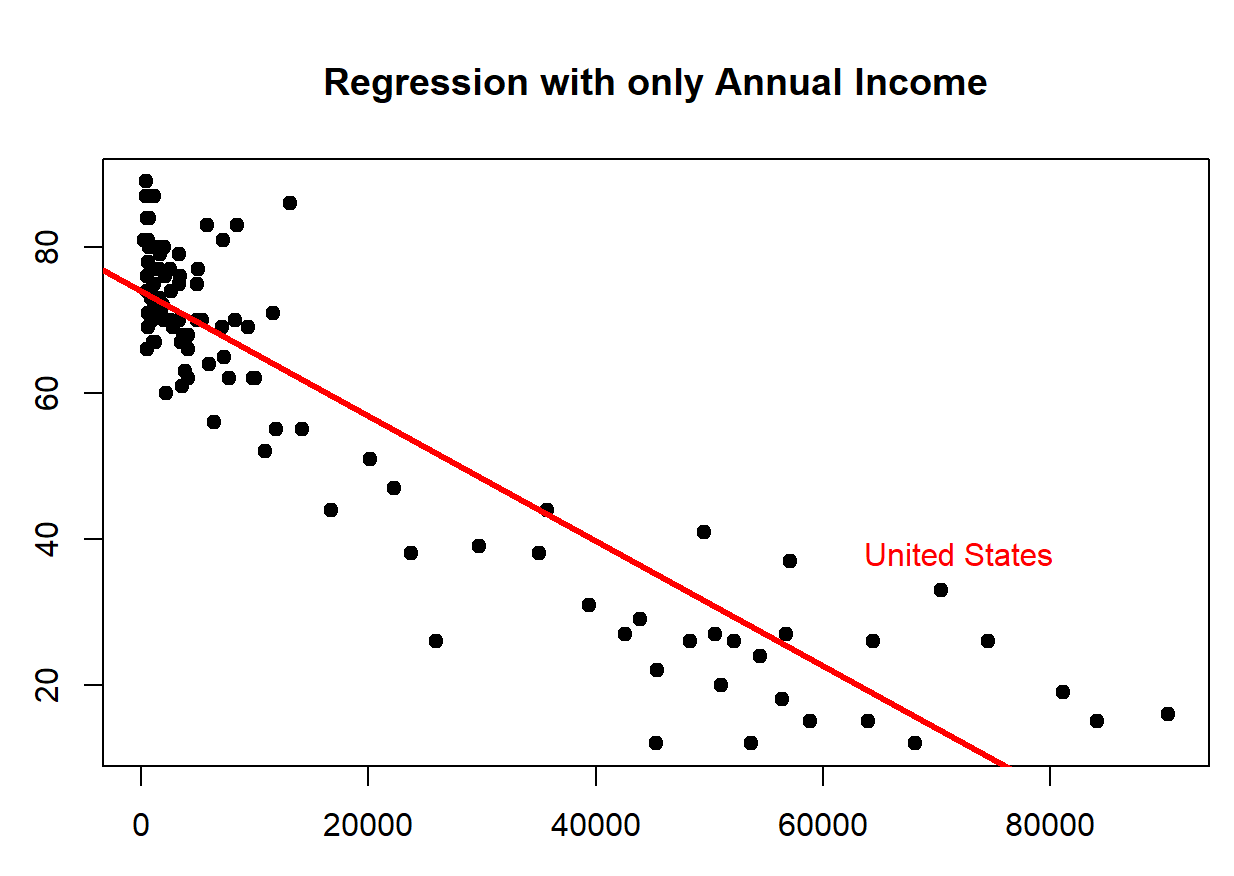
the corruption index (isolated variable)

```
singular_regression_full <- lm(corruption_index ~ annual_income, data = corruption_data)
summary(singular_regression_full)
```

```
##
## Call:
## lm(formula = corruption_index ~ annual_income, data = corruption_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.6891  -5.5876  -0.7898   6.1839  23.2769
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.392e+01  1.050e+00   70.43   <2e-16 ***
## annual_income -8.560e-04  3.591e-05  -23.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.901 on 108 degrees of freedom
## Multiple R-squared:  0.8403, Adjusted R-squared:  0.8388
## F-statistic: 568.2 on 1 and 108 DF,  p-value: < 2.2e-16
```

Regression Model with only one coefficient (summary shown above):

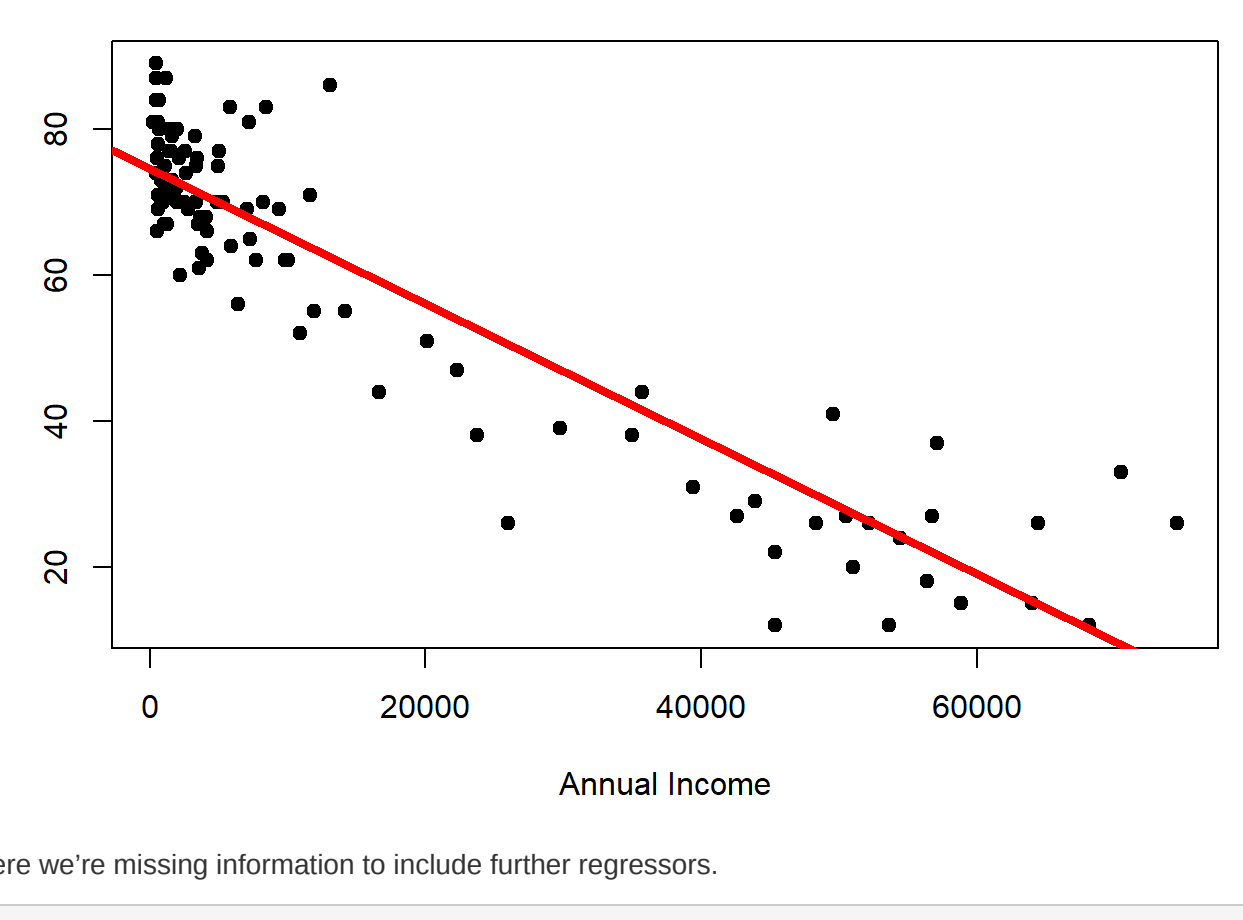
```
plot(corruption_data$annual_income, corruption_data$corruption_index, pch=19, xlab = "Annual Income", ylab = "Cor
ruption Index")
abline(singular_regression_full, col = "red", lwd = 3)
title("Regression with only Annual Income")
text(72000, 38, "United States", col = "red")
```



Note that when excluding income

over 80K, regression line becomes slightly steeper, since 3 countries with high average annual income and low corruption index were slightly skewing the regression line, namely Switzerland, Norway, and Luxembourg (ordered from most annual income to least).

```
corruption_data_excluding_income_over_80k <- filter(corruption_data, annual_income <= 80000)
singular_regression_excluding_income_over_80k <- lm (corruption_index ~ annual_income, data = corruption_data_exc
luding_income_over_80k)
plot(corruption_data_excluding_income_over_80k$annual_income, corruption_data_excluding_income_over_80k$corruptio
n_index, pch=19, xlab = "Annual Income", ylab = "Corruption Index")
abline(singular_regression_excluding_income_over_80k, col = "red", lwd = 4)
title("Regression Excluding Income over $80k")
```



Filter data to remove empty rows

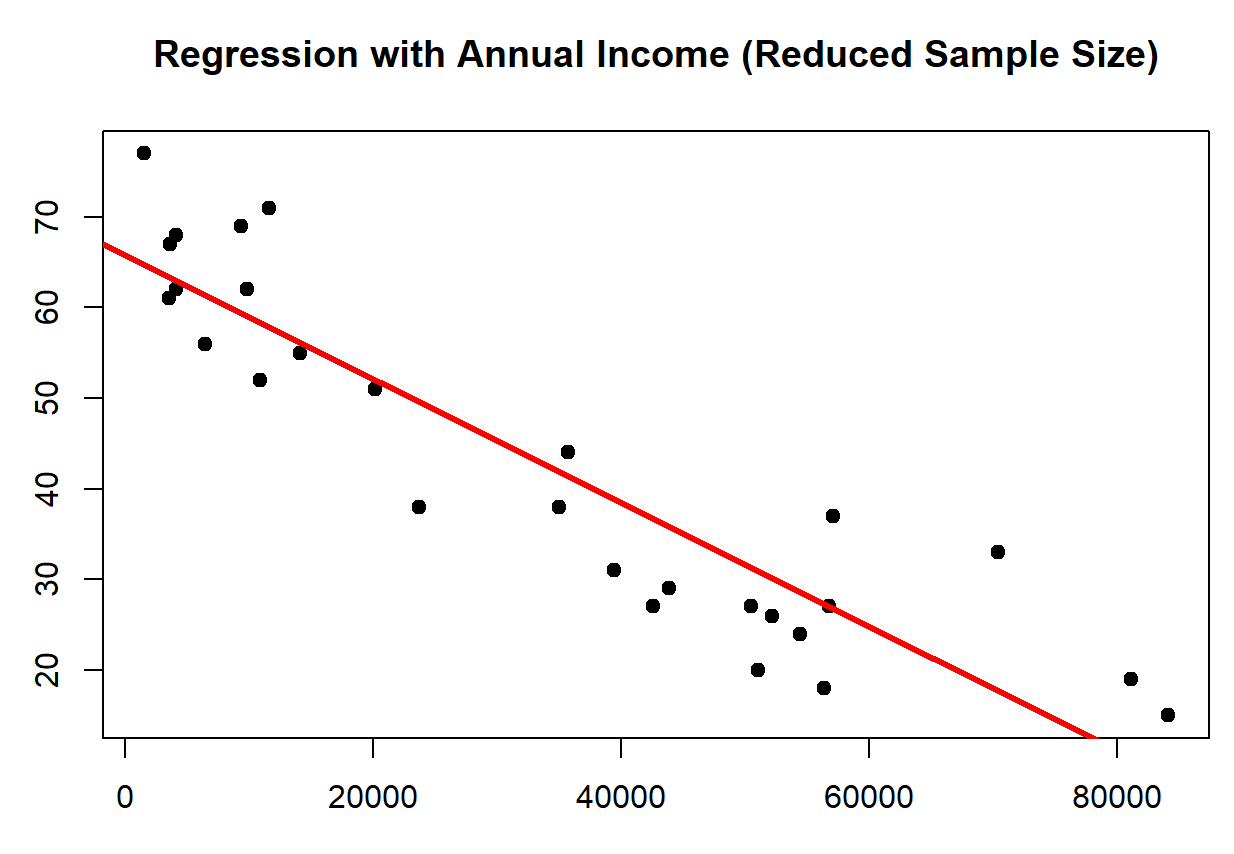
where we're missing information to include further regressors.

```
corruption_data_adjusted_for_missing_data <- filter(corruption_data, !is.na(tourists_in_millions) & !is.na(cost_i
ndex))
number_of_rows_in_new_dataset <- nrow(corruption_data_adjusted_for_missing_data)
print(number_of_rows_in_new_dataset)
```

```
## [1] 28
```

Running single regression with adjusted and reduced sample size (only 28 observations). Note how regression line becomes even steeper, mainly due to higher variability (not necessarily due to holding a stronger relationship)

```
single_regression_adjusted_for_missing_data <- lm (corruption_index ~ annual_income, data = corruption_data_adjus
ted_for_missing_data)
plot(corruption_data_adjusted_for_missing_data$annual_income, corruption_data_adjusted_for_missing_data$corruptio
n_index, pch=19, xlab = "Annual Income", ylab = "Corruption Index")
abline(single_regression_adjusted_for_missing_data, col = "red", lwd = 3)
title("Regression with Annual Income (Reduced Sample Size)")
```



Adding one more coefficient to the

regression: cost index (the higher cost index, the more expensive products and services are in that country relative to the US). This filtered data and comparison does not take tourist count into account, as it would significantly limit our available data set, from 110 observations all the way to 28.

```
corruption_data_without_cost_index_missing_data <- filter (corruption_data, !is.na(cost_index))
regr_w_one_regressor_excluding_tourism <- lm(corruption_index ~annual_income, data = corruption_data_without_cost
_index_missing_data)
regr_w_two_regressors_excluding_tourism <- lm(corruption_index ~ annual_income + cost_index, data = corruption_da
ta_without_cost_index_missing_data)

summary(regr_w_one_regressor_excluding_tourism)
```

```
##
## Call:
## lm(formula = corruption_index ~ annual_income, data = corruption_data_without_cost_index_missing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.3999  -5.6671   0.0193   6.2473  18.4017
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.973e+01  1.465e+00   47.60   <2e-16 ***
## annual_income -7.828e-04  4.041e-05  -19.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.879 on 69 degrees of freedom
## Multiple R-squared:  0.8447, Adjusted R-squared:  0.8424
## F-statistic: 375.3 on 1 and 69 DF,  p-value: < 2.2e-16
```

Note how there was only a very small change in R^2 when comparing the one regression model with the two regression model.

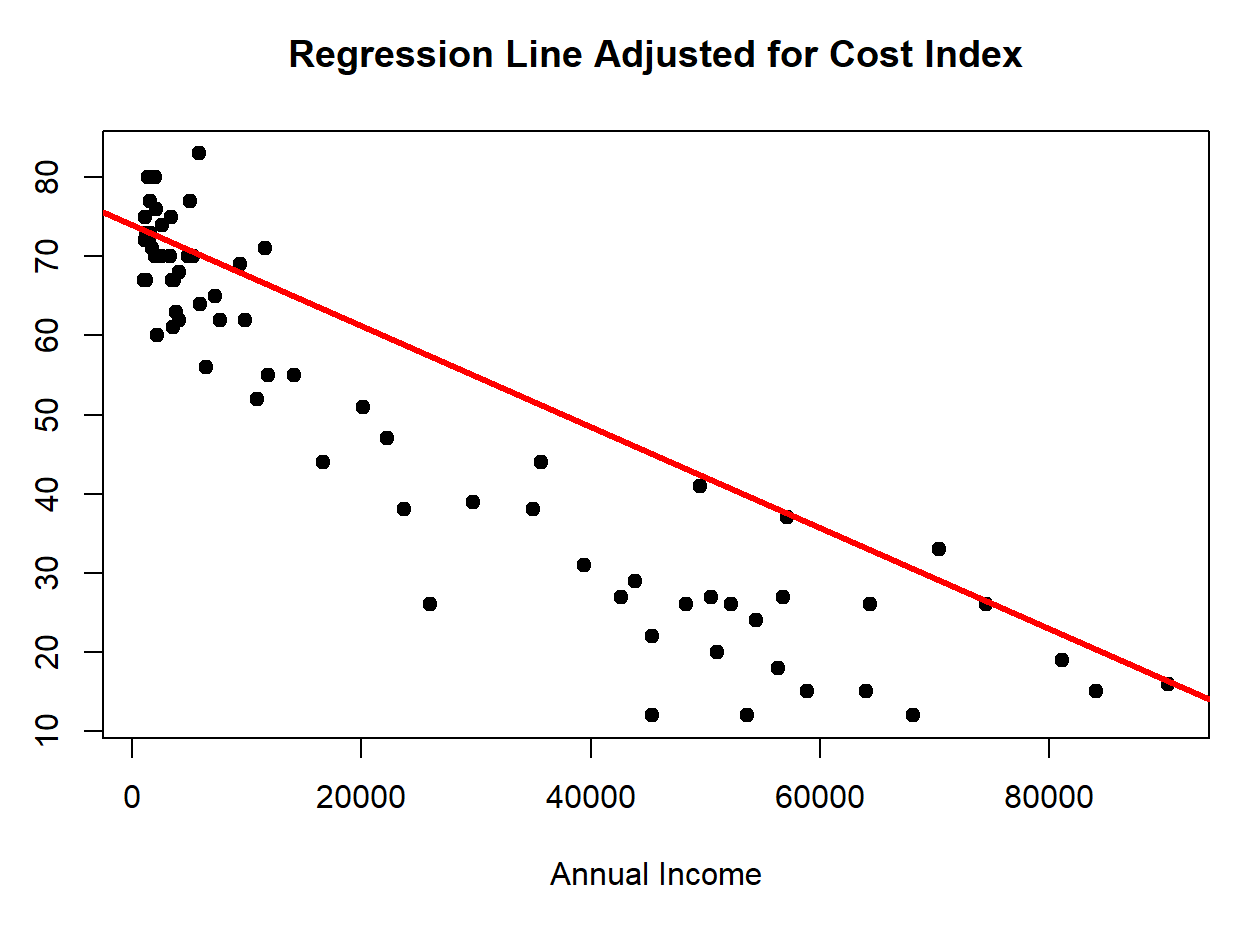
```
summary(regr_w_two_regressors_excluding_tourism)
```

```
##
## Call:
## lm(formula = corruption_index ~ annual_income + cost_index, data = corruption_data_without_cost_index_missing_
data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.3887  -5.8256  -0.9725   5.8159  19.5544
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   73.9358986  3.0962059  23.880  < 2e-16 ***
## annual_income -0.0006379  0.0001023  -6.235 3.29e-08 ***
## cost_index    -0.1213595  0.0788938  -1.538    0.129
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.793 on 68 degrees of freedom
## Multiple R-squared:  0.8499, Adjusted R-squared:  0.8455
## F-statistic: 192.5 on 2 and 68 DF,  p-value: < 2.2e-16
```

```
plot(corruption_data_without_cost_index_missing_data$annual_income, corruption_data_without_cost_index_missing_da
ta$corruption_index, pch=19, xlab = "Annual Income", ylab = "Corruption Index")
abline(regr_w_two_regressors_excluding_tourism, col = "red", lwd = 3)
```

```
## Warning in abline(regr_w_two_regressors_excluding_tourism, col = "red", : only
## using the first two of 3 regression coefficients
```

```
title("Regression Line Adjusted for Cost Index")
```



Tourists (in millions) may not be

statistically significant when running full regression. Recommend to remove it as a regressor and replace with another if more data were to become available on quantifying factors such as "political stability" or "safety" that is in alignment with data. Below is the regression with all 3 variables included.

```
multi_regr3 <- lm(corruption_index ~ annual_income + tourists_in_millions + cost_index, data = corruption_data_ad
justed_for_missing_data)
summary(multi_regr3)
```

```
##
## Call:
## lm(formula = corruption_index ~ annual_income + tourists_in_millions +
##     cost_index, data = corruption_data_adjusted_for_missing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.413  -4.615  -1.471   5.411  12.168
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          76.7902916  4.9018760  15.665  4.2e-14 ***
## annual_income        -0.0003525  0.0001396  -2.525   0.0186 *
## tourists_in_millions  0.0434778  0.0593272   0.733   0.4707
## cost_index           -0.3133511  0.1212110  -2.585   0.0162 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.26 on 24 degrees of freedom
## Multiple R-squared:  0.8706, Adjusted R-squared:  0.8545
## F-statistic: 53.84 on 3 and 24 DF,  p-value: 8.302e-11
```

Comparing only the 28 observations for all 3 coefficients with model 1 (just the annual income) and model 2 (with all 3 regressors)

```
library(texreg)
```

```
## Warning: package 'texreg' was built under R version 4.3.2
```

```
## Version:  1.39.3
## Date:     2023-11-09
## Author:   Philip Leifeld (University of Essex)
##
## Consider submitting praise using the praise or praise_interactive functions.
## Please cite the JSS article in your publications -- see citation("texreg").
```

```
screenreg(list(single_regression_adjusted_for_missing_data, multi_regr3))
```

```
##
## ==========================================
##                       Model 1    Model 2
## ------------------------------------------
## (Intercept)           65.79 ***  76.79 ***
##                       (2.49)     (4.90)
## annual_income         -0.00 ***  -0.00 *
##                       (0.00)     (0.00)
## tourists_in_millions              0.04
##                                  (0.06)
## cost_index                       -0.31 *
##                                  (0.12)
## ------------------------------------------
## R^2                    0.83       0.87
## Adj. R^2               0.83       0.85
## Num. obs.             28         28
## ==========================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Heteroskedastic F stat shown below for 3 regressors

```
linearHypothesis(multi_regr3, c("annual_income=0", "tourists_in_millions=0", "cost_index=0"), white.adjust = "hc
1")
```

	Res.Df	Df	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>
1	27	NA	NA	NA
2	24	3	54.64962	7.113285e-11

2 rows

Heteroskedastic F stat for two regressor model

```
linearHypothesis(regr_w_two_regressors_excluding_tourism, c("annual_income=0", "cost_index=0"), white.adjust = "h
c1")
```

	Res.Df	Df	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>
1	70	NA	NA	NA
2	68	2	146.8352	2.102151e-25

2 rows

Comparing 71 observations with model 1 (with one regressor) and model 2 (with annual income and cost index as the regressors)- since tourism might not be statistically significant. This could potentially be the best model (due to the high sample size- and it also omits limited tourism data (which isn't even statistically significant per the p value). It's regression line adjusted for cost index is shown above.

```
screenreg(list(regr_w_one_regressor_excluding_tourism, regr_w_two_regressors_excluding_tourism))
```

```
##
## ===================================
##                Model 1    Model 2
## -----------------------------------
## (Intercept)    69.73 ***  73.94 ***
##                (1.46)     (3.10)
## annual_income  -0.00 ***  -0.00 ***
##                (0.00)     (0.00)
## cost_index                -0.12
##                           (0.08)
## -----------------------------------
## R^2             0.84       0.85
## Adj. R^2        0.84       0.85
## Num. obs.      71         71
## ===================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```