Data Analysis On The Relation Between Income and Happiness

Justin Luong

3/7/2021

Abstract

There has been a long-standing debate about the ability for money to determine happiness. In my economic development class, I learned that wealth doesn't always lead to happier citizens of a country, so I wanted to see what components of income were key determinants of happiness. The purpose of this study was to analyze and determine the relation between different aspects of income and happiness using average satisfaction as the response variable and average income, median income, income inequality, and gross domestic product (GDP) as the predictor variables. This was an observational and analytical study using a Kaggle dataset of 111 countries. For the original model, I conducted exploratory data analysis in the form of boxplots, residual plots, Q-Q plots, and scale-location plots. These plots showed that the model assumptions were likely to be met, althought there were multiple potential outliers. I also conducted several tests (such as the Breusch-Pagan Test for constancy of variance, F-test for regression relation, and Kolmogorov-Smirnov Test for normality) to verify the model's assumptions were met. The estimated regression function for it was $\hat{Y} = 2.8241577 + 0.0002048X_1 - 0.0001775X_2 + 0.0239614X_3 + 2.1305654X_4$. All assumptions were met besides the constancy of error variance. Also, all the predictors besides GDP were not statistically significant. Thus, I checked if any predictor could be dropped using Mallow's Cp and R^2 statistics. It was shown that three predictor variables were ideal. Using variance inflation factors (VIF), I decided to drop the median income predictor variable from the model because it had the largest VIF.

With the new reduced model, the constancy of error variance assumption was still violated, so I created a correlation matrix and interaction plots to see if there was any interactivity. It was discovered that average income and GDP had an interaction effect. Thus, I created new model with the interaction terms. However, the the constancy of error variance assumption was still violated. This led me to try out a square root transformation on the response and predictor variables which ended up passing the all of the tests for the model assumptions. This was the final selected model. The estimated regression function for it was $\hat{Y} = 0.880053 + 0.018512X_1 + 0.063386X_2 + 0.809363X_3 - 0.012548X_1X_3$. This model suggests that average income, income inequality, and GDP are significant indicators of average satisfaction for countries, with GDP being the strongest indicator. It is also important that we subtract from the average satisfaction score in consideration of the interaction term of GDP and average income. This model shows that income does have an effect on average satisfaction level of countries, and that it is important to consider the many different aspects of income such as average income, income inequality, and GDP.

References

Below is the original happiness and income dataset from Kaggle.

https://www.kaggle.com/levyedgar44/income-and-happiness-correction

Data Analysis

Data Summary and Exploratory Analysis

```r
rm(list = ls())
setwd("C:/Users/justi/Documents/STA 108/Datasets/")
happiness_data <- read.csv('happyscore_income.csv', header = TRUE)

# I will analyze the effect on average satisfaction from average income, median income, income inequali
avg_satisfaction_lm <- lm(avg_satisfaction ~ avg_income + median_income + income_inequality + GDP, happ
summary(avg_satisfaction_lm)
```

```
##
## Call:
## lm(formula = avg_satisfaction ~ avg_income + median_income +
##     income_inequality + GDP, data = happiness_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11336 -0.44302  0.07791  0.47025  1.82826
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.8241577  0.7218208   3.913 0.000162 ***
## avg_income         0.0002048  0.0002224   0.921 0.359339
## median_income     -0.0001775  0.0002583  -0.687 0.493430
## income_inequality  0.0239614  0.0169986   1.410 0.161582
## GDP                2.1305654  0.3700061   5.758 8.38e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8168 on 106 degrees of freedom
## Multiple R-squared:  0.6506, Adjusted R-squared:  0.6374
## F-statistic: 49.34 on 4 and 106 DF,  p-value: < 2.2e-16
```

For average satisfaction, it seems that GDP is the only statistically significant predictor variable based on the comparison of p-values and the significance levels. Now, I will conduct exploratory data analysis and tests to assess the model's assumptions before attempting to find a better fitting model.

Interpretation of Model

From the regression summary, we see that the estimated regression function is $\hat{Y} = 2.8241577 + 0.0002048X_1 - 0.0001775X_2 + 0.0239614X_3 + 2.1305654X_4$.

The interpretation of b_0 is that 2.8241577 is the mean value that we would predict for the average satisfaction level if $X_1 = X_2 = X_3 = X_4 = 0$. It is also known as the y-intercept.

The interpretation of b_1 is that with a one unit increase in X_1 (average income), there will be an increase of 0.0002048 to the average satisfaction level while holding X_2, X_3, and X_4 constant.

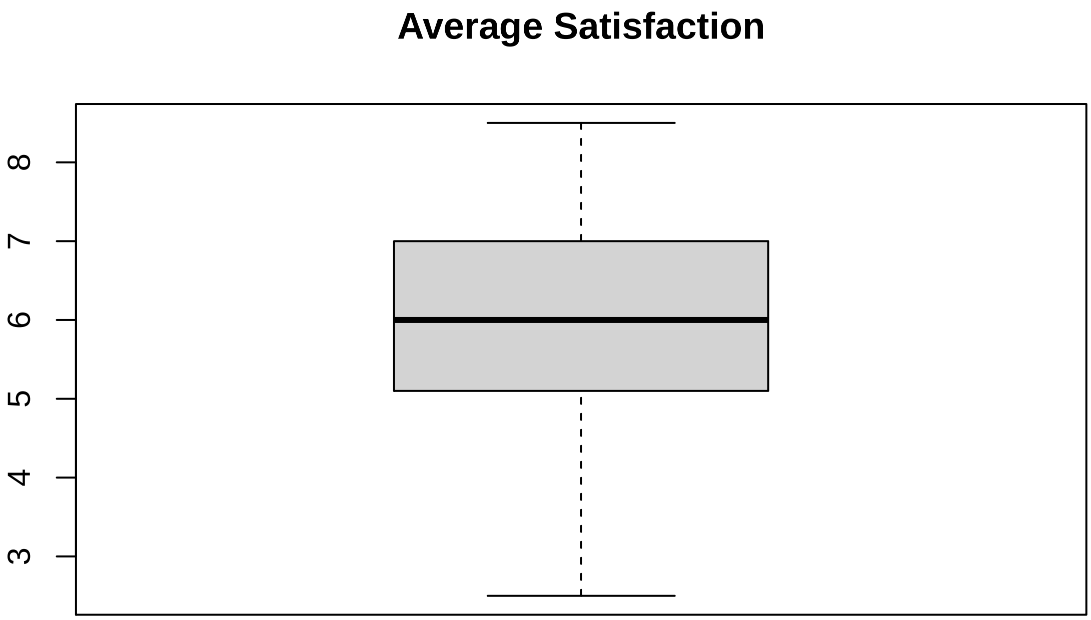
The interpretation of b_2 is that with a one unit increase in X_2 (median income), there will be a decrease of 0.0001775 to the average satisfaction level while holding X_1, X_3, and X_4 constant.

The interpretation of b_3 is that with a one unit increase in X_3 (income inequality), there will be an increase of 0.0239614 to the average satisfaction level while holding X_1, X_2, and X_4 constant.

The interpretation of b_4 is that with a one unit increase in X_4 (GDP), there will be an increase of 2.1305654 to the average satisfaction level while holding X_1, X_2, and X_3 constant.

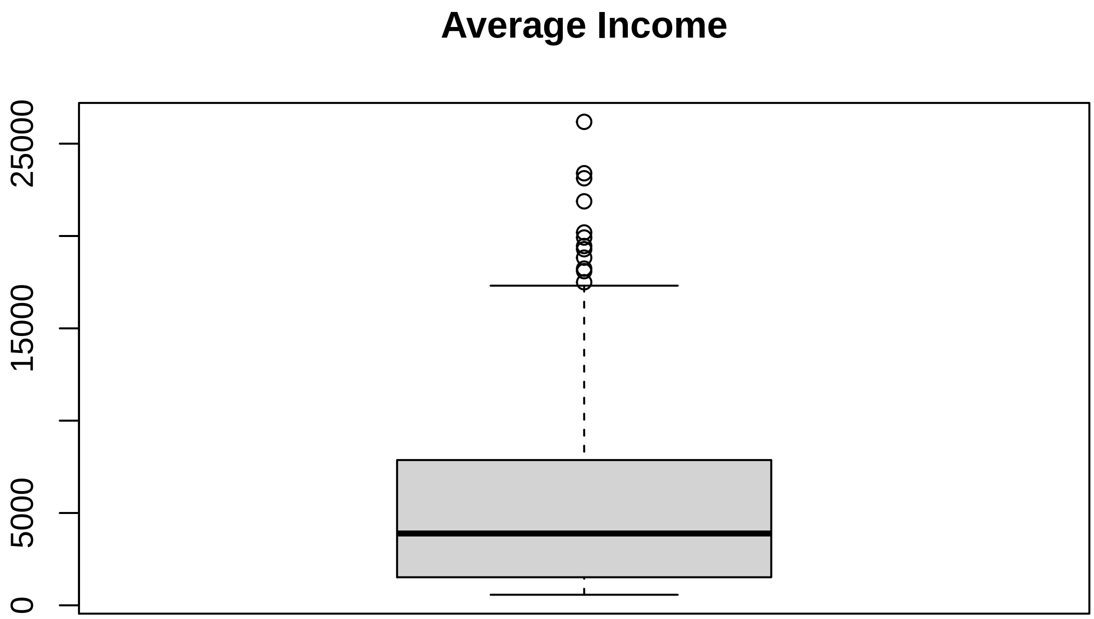
Boxplots

```
boxplot(happiness_data$avg_satisfaction, main = "Average Satisfaction")
```



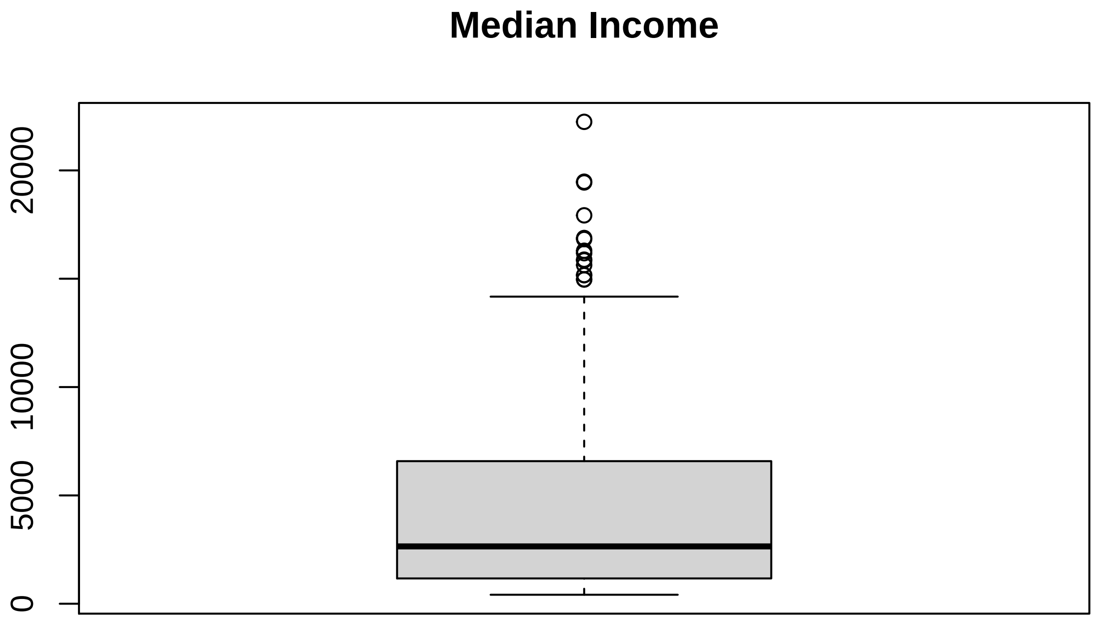
The boxplot for average satisfaction shows a normal distribution and no outliers.

```
boxplot(happiness_data$avg_income, main = "Average Income")
```



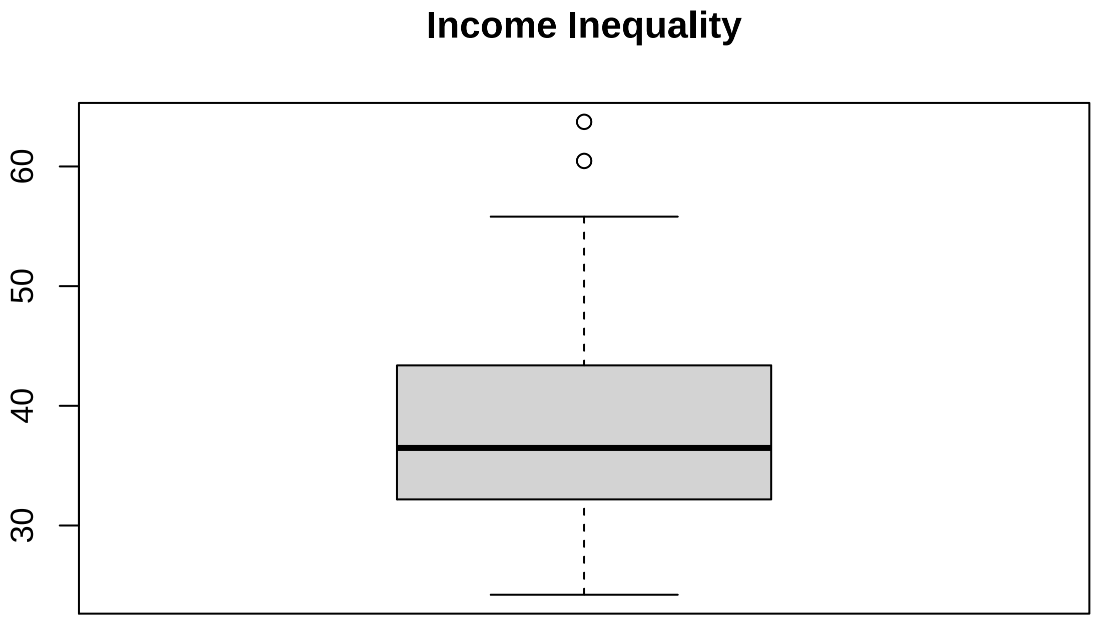
The boxplot for average income shows that there is a right-skewed distribution as well as many outliers in the upper extreme section.

```
boxplot(happiness_data$median_income, main = "Median Income")
```



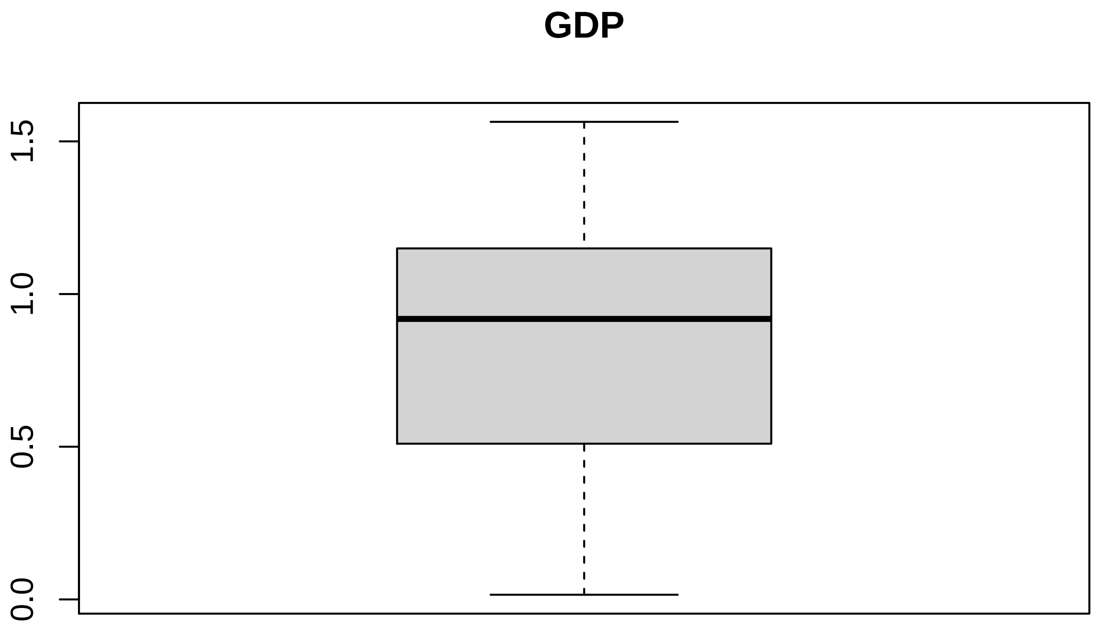
The boxplot for median income shows that there is a right-skewed distribution as well as many outliers in the upper extreme section.

```r
boxplot(happiness_data$income_inequality, main = "Income Inequality")
```



The boxplot for income inequality shows that there is slightly right-skewed distribution with two outliers in the upper extreme section.

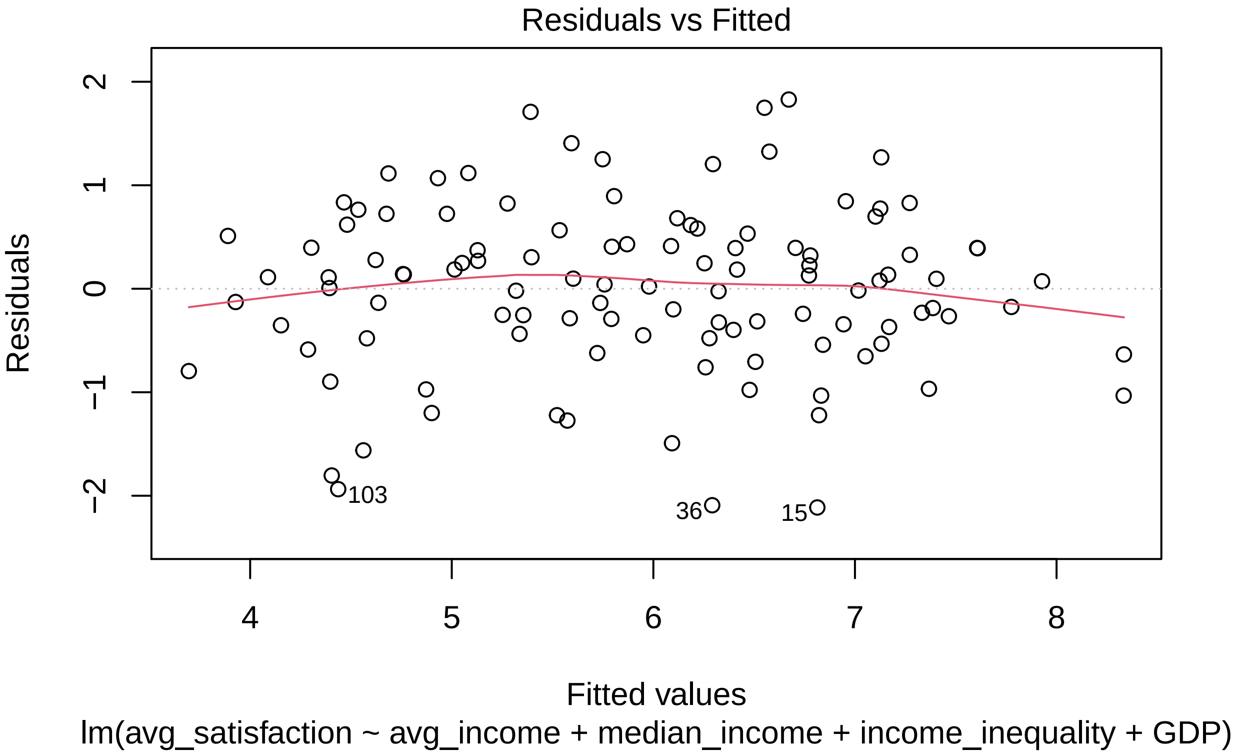
```r
boxplot(happiness_data$GDP, main = "GDP")
```

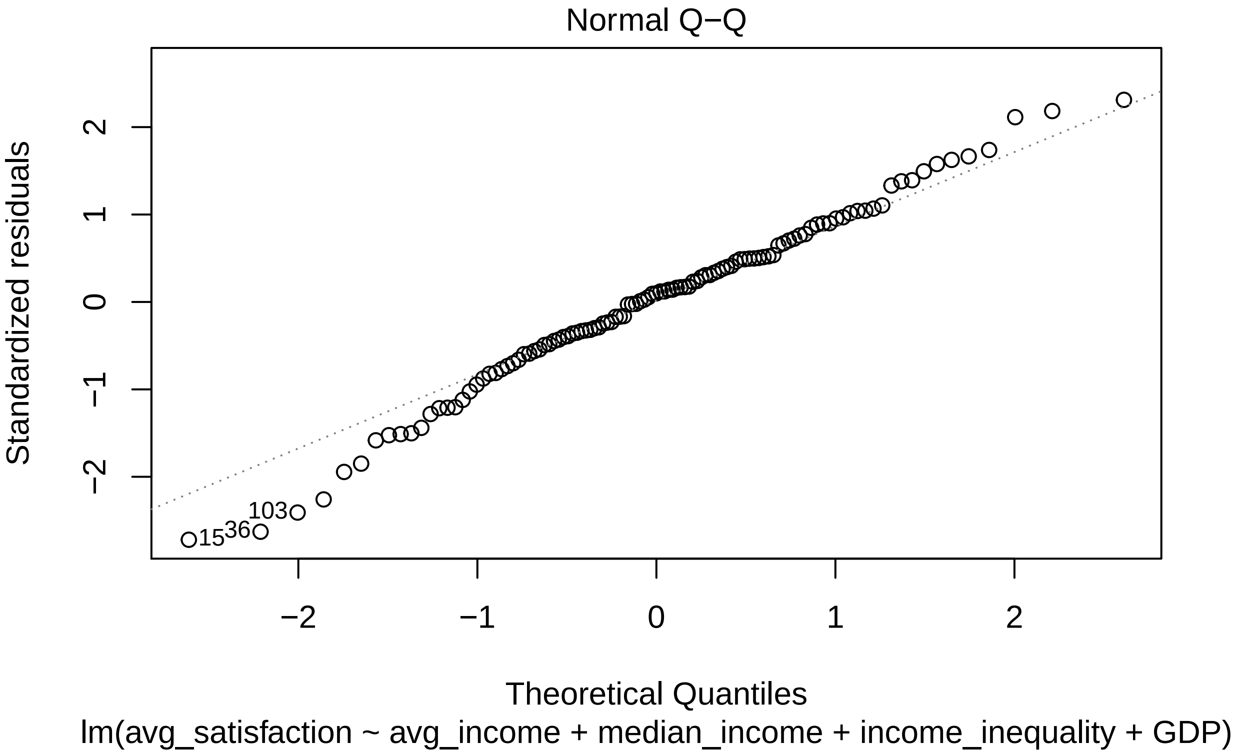


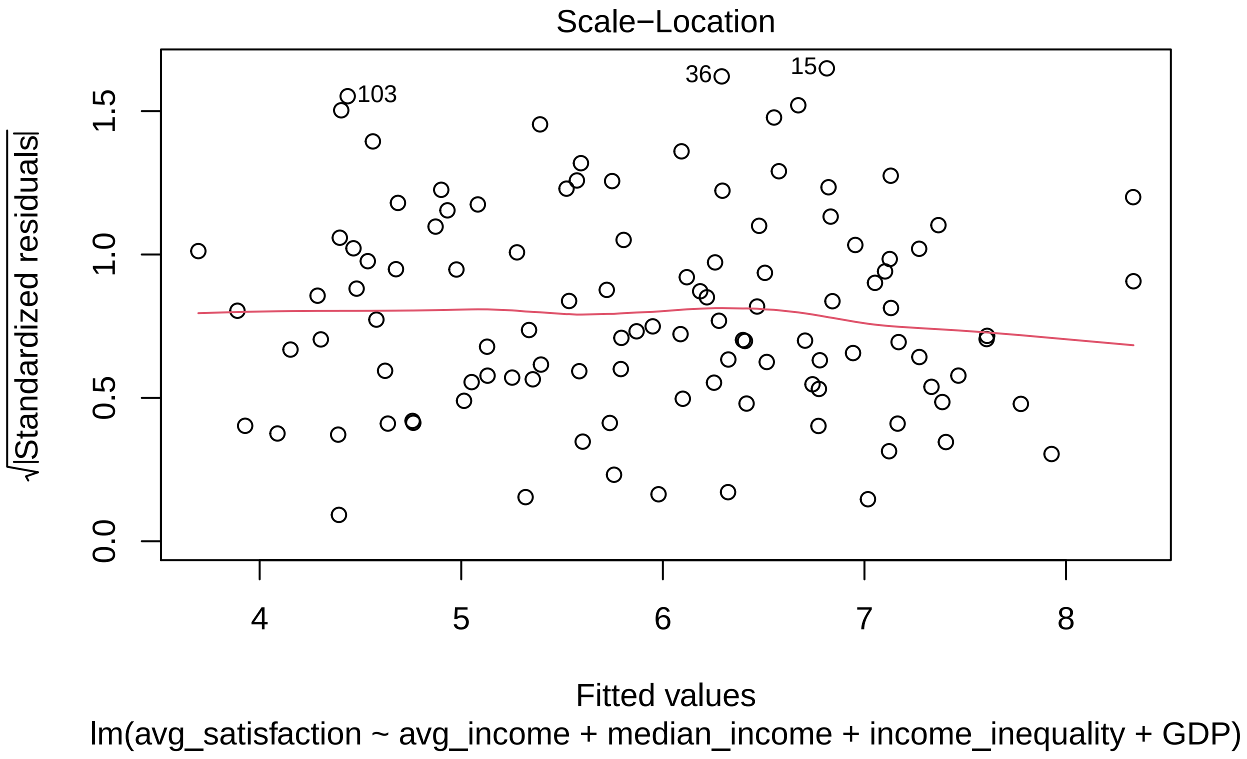
The boxplot for GDP shows that there is a normal distrbution and no outliers.

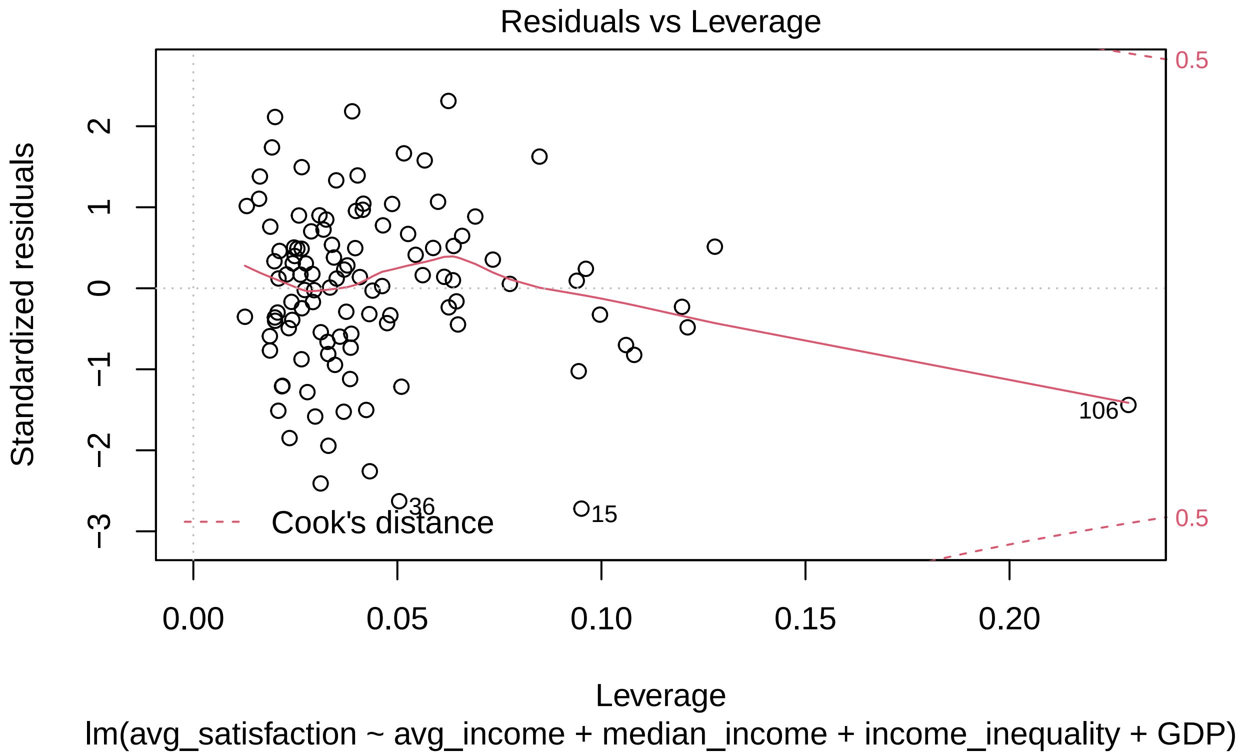
Model Diagnostics

```
plot(avg_satisfaction_lm)
```


Residuals vs Fitted
Residuals
Fitted values
lm(avg_satisfaction ~ avg_income + median_income + income_inequality + GDP)
2
1
0
−1
−2
4
5
6
7
8
103
36
15


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
15
36
103
−2
−1
0
1
2
Theoretical Quantiles
lm(avg_satisfaction ~ avg_income + median_income + income_inequality + GDP)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
4
5
6
7
8
Fitted values
lm(avg_satisfaction ~ avg_income + median_income + income_inequality + GDP)
103
36
15



Although there are some outliers, there is a fairly horizontal line in the Residuals vs. Fitted plot without any distinct patterns which indicates a linear relationship. In the Normal Q-Q plot, the residuals are follow the Q-Q line closely for the most part aside from deviation on the ends, which indicates that the normality assumption is reasonable. The Scale-Location plot has a horizontal line with spread out points which suggests that our assumption of constant error variance is reasonably met. The Residuals vs. Leverage plot does not show any points with a large Cook's distance, which indicates that there are no highly influential cases in our model.

Breusch-Pagan Test For Constancy of Error Variance

I will conduct the test with a significance level of 0.05.

Hypotheses: H_0 = Error variance is constant vs. H_a = Error variance is not constant

Decision Rule:

$\chi^2_{BP} \leq \chi^2_{(1-\alpha;p-1)}$ = Conclude H_0

$\chi^2_{BP} > \chi^2_{(1-\alpha;p-1)}$ = Conclude H_a

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
bptest(avg_satisfaction_lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  avg_satisfaction_lm
## BP = 12.754, df = 4, p-value = 0.01254
```

```r
qchisq(1-0.05,4)
```

```
## [1] 9.487729
```

Conclusion

Since $12.754 > 9.487729$, we conclude H_a that the error variance is not constant.

Overall F-Test For Regression Relation

I will conduct the test with a significance level of 0.05.

Hypotheses

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_a : not all $\beta_k (k = 1, ..., p - 1)$ equal zero

Decision Rule

$F^* \leq F(1 - \alpha; p - 1, n - p)$, conclude H_0

$F^* > F(1 - \alpha; p - 1, n - p)$, conclude H_a

```r
# ANOVA Table
anova(avg_satisfaction_lm)
```

```
## Analysis of Variance Table
##
## Response: avg_satisfaction
##                    Df Sum Sq Mean Sq  F value    Pr(>F)
## avg_income          1 96.085  96.085 144.0376 < 2.2e-16 ***
## median_income       1 13.438  13.438  20.1450 1.831e-05 ***
## income_inequality   1  0.025   0.025   0.0379    0.8459
## GDP                 1 22.118  22.118  33.1567 8.379e-08 ***
## Residuals         106 70.711   0.667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 0.95 quantile of F-distribution with df = 4, 106
qf(1-0.05, 4, 106)
```

```
## [1] 2.45738
```

```
# F* = MSR/MSE
((96.085 + 13.438 + 0.025 + 22.1118)/4) / 0.667
```

```
## [1] 49.34775
```

Conclusion

Since $49.34775 > 2.45738$, we conclude H_a that not all $\beta_k (k = 1, ..., p - 1)$ equal zero. At the $\alpha = 5$ significance level, there is sufficient evidence that the average satisfaction level is related to the average income, median income, income inequality, and GDP.

KS Test for Normality Assumption

I will use a significance level of 0.05 for this KS test.

Hypotheses

H_0: normality assumption holds

H_a normality assumption does not hold

Decision Rule

If p-value $> \alpha$: Conclude H_0

If p-value $\leq \alpha$: Conclude H_a

```
set.seed(100)
y.mean <- mean(avg_satisfaction_lm$fitted.values)
y.std <- sd(avg_satisfaction_lm$fitted.values)
ks.test((avg_satisfaction_lm$fitted.values - y.mean)/y.std, rnorm(111))
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  (avg_satisfaction_lm$fitted.values - y.mean)/y.std and rnorm(111)
## D = 0.13514, p-value = 0.2629
## alternative hypothesis: two-sided
```

Conclusion

Since $0.2629 > 0.05$, we cannot reject the null hypothesis and thus conclude that our normality assumption holds.

Model Selection

I will try to find a better model because the regression summary showed that GDP was the only statistically significant predictor variable and the constancy of error variance assumption was violated.

```
library(tidyverse)
```

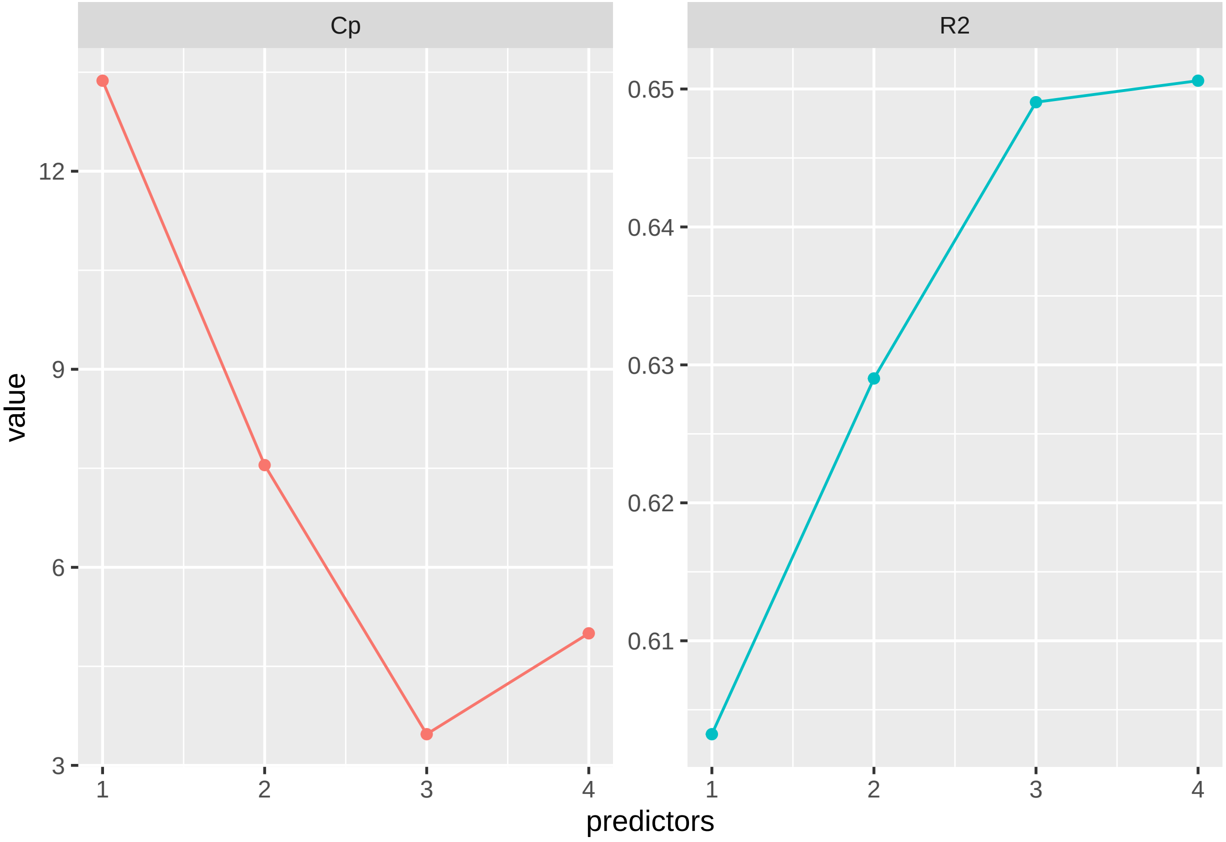
```
## Warning: package 'tidyverse' was built under R version 4.0.4
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.0.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.0.4
```

```
adj_happiness_data <- read.csv('happiness_income.csv', header = TRUE)
best_subset <- regsubsets(avg_satisfaction ~ ., adj_happiness_data, nvmax = 4)
results <- summary(best_subset)
tibble(predictors = 1:4,
       R2 = results$rsq,
       Cp = results$cp) %>%
  gather(statistic, value, -predictors) %>%
  ggplot(aes(predictors, value, color = statistic)) +
  geom_line(show.legend = F) +
  geom_point(show.legend = F) +
  facet_wrap(~ statistic, scales = "free")
```



From the plot of the Mallow's Cp values and the amount of predictor variables, it is apparent that having 3 predictors variables leads to the smallest Cp value.

From the plot of the R^2 and amount of predictor variables, it is apparent that having 3 predictor variable is ideal as that is when the curve starts to flatten out.

Therefore, I will select a model with 3 predictor variables.

Variance Inflation Factor

We will drop the variable with the highest variance inflation factor.

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.4
```

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

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

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

```
## The following object is masked from 'package:purrr':
##
##     some
```

```r
vif(avg_satisfaction_lm)
```

```
##        avg_income     median_income income_inequality               GDP
##        342.472668        343.513203          3.342444          3.391371
```

Given that the predictor variable median income has the biggest variance inflation factor, I will try out a model with 3 predictor variables where median income is excluded. In addition, it is interesting that a higher median income actually reduces the average satisfaction when I expected it to increase. This also leads me to believe it should be removed from the model.

Reduced Model: 3 Predictor Variables

The reduced model we will choose is $\hat{Y}_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$, X_1 = Average Income, X_2 = Income Inequality, and X_3 = GDP.

```r
new_happiness_data <- subset(adj_happiness_data, select = -c(median_income))
new_happiness_lm <- lm(avg_satisfaction ~ ., data = new_happiness_data)
summary(new_happiness_lm)
```

```
##
## Call:
## lm(formula = avg_satisfaction ~ ., data = new_happiness_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15374 -0.43145  0.09401  0.44225  1.93223
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.446e+00  4.662e-01   5.247 7.87e-07 ***
## avg_income        5.261e-05  2.129e-05   2.471  0.01505 *
## income_inequality 3.338e-02  1.004e-02   3.324  0.00121 **
## GDP               2.221e+00  3.451e-01   6.435 3.57e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.8147 on 107 degrees of freedom
## Multiple R-squared:  0.649,  Adjusted R-squared:  0.6392
## F-statistic: 65.96 on 3 and 107 DF,  p-value: < 2.2e-16
```

It is a good sign that all the p-values for all three predictors are below the significance level. I will conduct test the model assumptions to further verify the good fit.

Breusch-Pagan Test For Constancy of Error Variance

We will conduct the test with a significance level of 0.05.

Hypotheses: H_0 = Error variance is constant vs. H_a = Error variance is not constant

Decision Rule:

$\chi^2_{BP} \leq \chi^2_{(1-\alpha;p-1)}$ = Conclude H_0

$\chi^2_{BP} > \chi^2_{(1-\alpha;p-1)}$ = Conclude H_a

```r
bptest(new_happiness_lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  new_happiness_lm
## BP = 13.157, df = 3, p-value = 0.004309
```

```r
qchisq(1-0.05,3)
```

```
## [1] 7.814728
```

Conclusion

Since $13.157 > 7.814728$, we conclude H_a that the error variance is not constant.

Overall F-Test For Regression Relation

I will conduct the test with a significance level of 0.05.

Hypotheses

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_a : not all $\beta_k (k = 1, ..., p - 1)$ equal zero

Decision Rule

$F^* \leq F(1 - \alpha; p - 1, n - p)$, conclude H_0

$F^* > F(1 - \alpha; p - 1, n - p)$, conclude H_a

```
anova(new_happiness_lm)
```

```
## Analysis of Variance Table
##
## Response: avg_satisfaction
##                    Df Sum Sq Mean Sq F value    Pr(>F)
## avg_income          1 96.085  96.085 144.751 < 2.2e-16 ***
## income_inequality   1  7.780   7.780  11.720 0.0008773 ***
## GDP                 1 27.487  27.487  41.409 3.569e-09 ***
## Residuals         107 71.026   0.664
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
# F* = MSR/MSE
((96.085 + 7.780 + 27.487)/3) / 0.664
```

```
## [1] 65.93976
```

```
# 0.95 quantile of F-distribution with df = 3, 107
qf(1-0.05, 3, 107)
```

```
## [1] 2.68949
```

Conclusion

Since $65.93976 > 2.68949$, we conclude H_a that not all $\beta_k (k = 1, ..., p - 1)$ equal zero. At the $\alpha = 5$ significance level, there is sufficient evidence that the average satisfaction level is related to the average income, income inequality, and GDP.

KS Test For Normality Assumption

I will use a significance level of 0.05 for this KS test.

Hypotheses

H_0: normality assumption holds

H_a normality assumption does not hold

Decision Rule

If p-value $> \alpha$: Conclude H_0

If p-value $\leq \alpha$: Conclude H_a

```
set.seed(100)
new.y.mean <- mean(new_happiness_lm$fitted.values)
new.y.std <- sd(new_happiness_lm$fitted.values)
ks.test((new_happiness_lm$fitted.values - new.y.mean)/new.y.std, rnorm(111))
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  (new_happiness_lm$fitted.values - new.y.mean)/new.y.std and rnorm(111)
## D = 0.14414, p-value = 0.1991
## alternative hypothesis: two-sided
```

Conclusion

Since $0.1991 > 0.05$, we conclude H_0 that the normality assumption holds.

Correlation Matrix

Given that these predictor variables are all related to income in some manner, we will check for correlation and interaction effects between the variables.

```
cor(new_happiness_data)
```

```
##                     avg_satisfaction avg_income income_inequality        GDP
## avg_satisfaction          1.00000000  0.6890431       -0.08247104  0.7766786
## avg_income                0.68904312  1.0000000       -0.38258727  0.8140243
## income_inequality        -0.08247104 -0.3825873        1.00000000 -0.3032042
## GDP                       0.77667857  0.8140243       -0.30320418  1.0000000
```

We will check for interaction effects between b_1 and b_3 because of their high correlation.

Interaction Effects

```
library(sjPlot)
```

```
## Warning: package 'sjPlot' was built under R version 4.0.4
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## Install package "strengejacke" from GitHub (`devtools::install_github("strengejacke/strengejacke")`)
```

```
library(sjmisc)
```

```
## Warning: package 'sjmisc' was built under R version 4.0.4
```

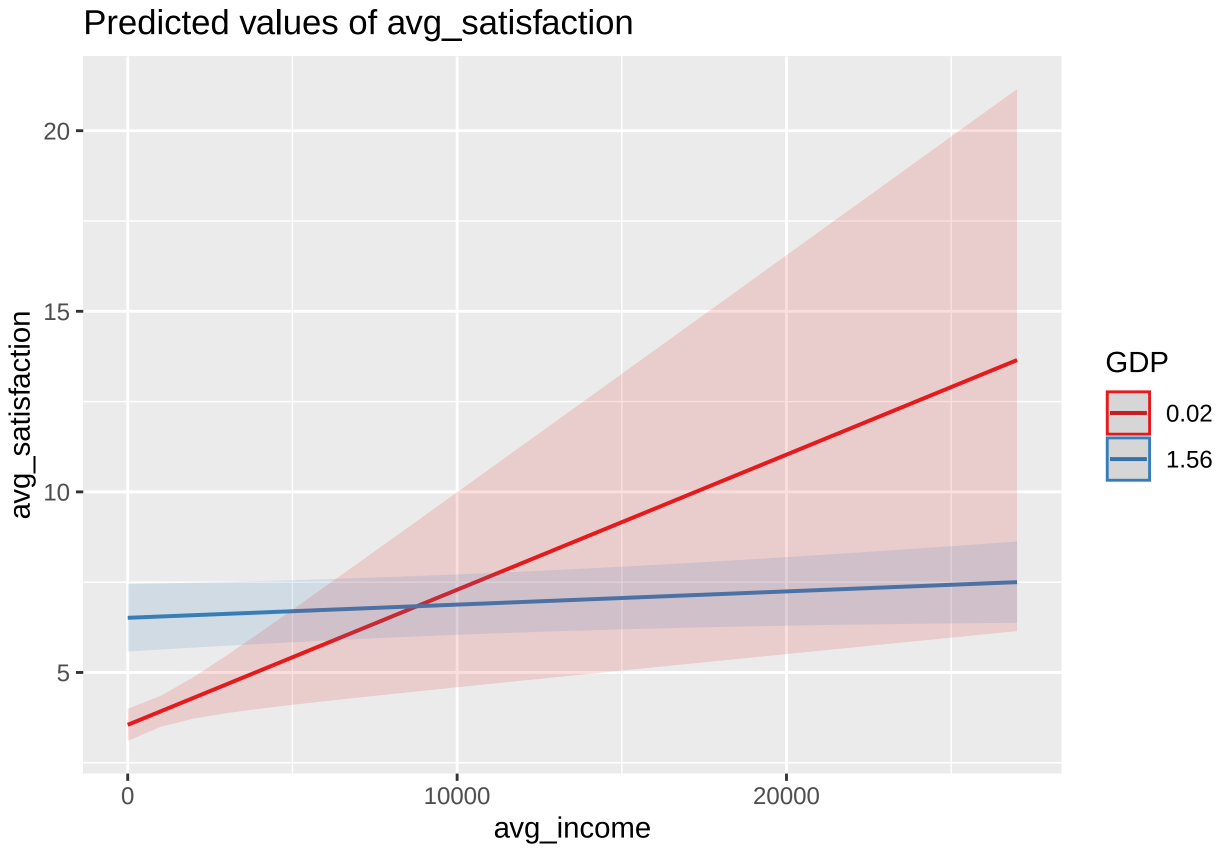
```
##
## Attaching package: 'sjmisc'
```

```
## The following object is masked from 'package:purrr':
##
##     is_empty
```

```
## The following object is masked from 'package:tidyr':
##
##     replace_na
```

```
## The following object is masked from 'package:tibble':
##
##     add_case
```

```
# beta1 * beta_3 interaction plot
b1b3_lm <- lm(avg_satisfaction ~ income_inequality + avg_income * GDP, new_happiness_data)
plot_model(b1b3_lm, type = "int", terms = c("avg_income", "GDP"))
```



Since the lines of the interaction plot intersect, it is apparent that interactivity exists between average income and GDP.

Interaction Model

I will add the $\beta_1\beta_3$ interaction term to the reduced model.

```
interaction_lm <- lm(avg_satisfaction ~ avg_income + income_inequality + GDP + avg_income * GDP, new_hap
summary(interaction_lm)
```

```
##
## Call:
## lm(formula = avg_satisfaction ~ avg_income + income_inequality +
##     GDP + avg_income * GDP, data = new_happiness_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96755 -0.46520  0.09662  0.53624  1.85506
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.319e+00  4.611e-01   5.028 2.03e-06 ***
## avg_income          3.785e-04  1.467e-04   2.580  0.01126 *
## income_inequality   3.105e-02  9.910e-03   3.133  0.00224 **
## GDP                 1.924e+00  3.637e-01   5.291 6.61e-07 ***
## avg_income:GDP     -2.191e-04  9.767e-05  -2.244  0.02692 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7998 on 106 degrees of freedom
## Multiple R-squared:  0.665,  Adjusted R-squared:  0.6523
## F-statistic: 52.59 on 4 and 106 DF,  p-value: < 2.2e-16
```

It is a good sign that all the p-values for all three predictors are below the significance level. I will conduct test the model assumptions to further verify the good fit.

Breusch-Pagan Test For Constancy of Error Variance We will conduct the test with a significance level of 0.05.

Hypotheses: H_0 = Error variance is constant vs. H_a = Error variance is not constant

Decision Rule:

$\chi^2_{BP} \leq \chi^2_{(1-\alpha;p-1)}$ = Conclude H_0

$\chi^2_{BP} > \chi^2_{(1-\alpha;p-1)}$ = Conclude H_a

```
bptest(interaction_lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  interaction_lm
## BP = 11.206, df = 4, p-value = 0.02434
```

```
qchisq(1-0.05,4)
```

```
## [1] 9.487729
```

Conclusion

Since 11.206 > 9.487729, we conclude H_a that the error variance is not constant. On the bright side, the marginal difference has gotten smaller.

Square Root Transformation

I will apply a square root transformation on the predictor variables to attempt to fix the lack of constancy of error variance.

```
sqrt_response <- sqrt(new_happiness_data$avg_satisfaction)

# Assign predictor variables to a new variable to make a new data frame
avg_income <- sqrt(new_happiness_data$avg_income)
income_inequality <- sqrt(new_happiness_data$income_inequality)
GDP <- sqrt(new_happiness_data$GDP)

# New data frame with square root transformation on response variable
sqrt_data <- data.frame(sqrt_response, avg_income, income_inequality, GDP)

# Transformed MLR
sqrt_lm <- lm(sqrt_response ~ avg_income + income_inequality + GDP + avg_income * GDP, sqrt_data)
summary(sqrt_lm)
```

```
##
## Call:
## lm(formula = sqrt_response ~ avg_income + income_inequality +
##     GDP + avg_income * GDP, data = sqrt_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48951 -0.08368  0.02394  0.11112  0.36390
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.880053   0.226363   3.888 0.000177 ***
## avg_income         0.018512   0.006738   2.747 0.007061 **
## income_inequality  0.063386   0.026718   2.372 0.019478 *
## GDP                0.809363   0.141640   5.714 1.02e-07 ***
## avg_income:GDP    -0.012548   0.005349  -2.346 0.020833 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1678 on 106 degrees of freedom
## Multiple R-squared:  0.678,  Adjusted R-squared:  0.6658
## F-statistic:  55.8 on 4 and 106 DF,  p-value: < 2.2e-16
```

It is a good sign that all the p-values for all three predictors are below the significance level. I will conduct test the model assumptions to further verify the good fit.

Breusch-Pagan Test For Constancy of Error Variance

We will conduct the test with a significance level of 0.01.

Hypotheses: H_0 = Error variance is constant vs. H_a = Error variance is not constant

Decision Rule:

$\chi^2_{BP} \leq \chi^2_{(1-\alpha;p-1)}$ = Conclude H_0

$\chi^2_{BP} > \chi^2_{(1-\alpha;p-1)}$ = Conclude H_a

```r
bptest(sqrt_lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  sqrt_lm
## BP = 10.592, df = 4, p-value = 0.03155
```

```r
qchisq(1-0.01,4)
```

```
## [1] 13.2767
```

Since $10.592 < 13.2767$, we conclude H_0 that the error variance is constant.

Overall F-Test For Regression Relation

I will conduct the test with a significance level of 0.05.

Hypotheses

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_a : not all $\beta_k (k = 1, ..., p - 1)$ equal zero

Decision Rule

$F^* \leq F(1 - \alpha; p - 1, n - p)$, conclude H_0

$F^* > F(1 - \alpha; p - 1, n - p)$, conclude H_a

```r
anova(sqrt_lm)
```

```
## Analysis of Variance Table
##
## Response: sqrt_response
##                    Df Sum Sq Mean Sq F value    Pr(>F)
## avg_income          1 5.0197  5.0197 178.230 < 2.2e-16 ***
## income_inequality   1 0.3171  0.3171  11.261   0.00110 **
## GDP                 1 0.7939  0.7939  28.188 6.101e-07 ***
## avg_income:GDP      1 0.1550  0.1550   5.504   0.02083 *
## Residuals         106 2.9854  0.0282
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
# F* = MSR/MSE
((5.0197 + 0.3171 + 0.7939 + 0.1550)/4) / 0.0282
```

```
## [1] 55.72429
```

```r
# 0.95 quantile of F-distribution with df = 4, 106
qf(1-0.05, 4, 106)
```

```
## [1] 2.45738
```

Since $55.72429 > 2.45738$, we conclude H_a that not all $\beta_k (k = 1, ..., p-1)$ equal zero. At the $\alpha = 5$ significance level, there is sufficient evidence that the average satisfaction level is related to the average income, income inequality, GDP, and interaction term of average income and GDP.

KS Test For Normality Assumption

I will use a significance level of 0.05 for this KS test.

Hypotheses

H_0: normality assumption holds

H_a normality assumption does not hold

Decision Rule

If p-value $> \alpha$: Conclude H_0

If p-value $\leq \alpha$: Conclude H_a

```r
set.seed(100)
sqrt.y.mean <- mean(sqrt_lm$fitted.values)
sqrt.y.std <- sd(sqrt_lm$fitted.values)
ks.test((sqrt_lm$fitted.values - sqrt.y.mean)/sqrt.y.std, rnorm(111))
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  (sqrt_lm$fitted.values - sqrt.y.mean)/sqrt.y.std and rnorm(111)
## D = 0.18018, p-value = 0.05445
## alternative hypothesis: two-sided
```

Since $0.05445 > 0.05$, we conclude H_0 that the normality assumption holds.

```r
summary(sqrt_lm)
```

```
##
## Call:
## lm(formula = sqrt_response ~ avg_income + income_inequality +
##     GDP + avg_income * GDP, data = sqrt_data)
##
## Residuals:
##       Min       1Q   Median       3Q      Max
## -0.48951 -0.08368  0.02394  0.11112  0.36390
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.880053   0.226363   3.888 0.000177 ***
## avg_income         0.018512   0.006738   2.747 0.007061 **
## income_inequality  0.063386   0.026718   2.372 0.019478 *
## GDP                0.809363   0.141640   5.714 1.02e-07 ***
## avg_income:GDP    -0.012548   0.005349  -2.346 0.020833 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1678 on 106 degrees of freedom
## Multiple R-squared:  0.678,  Adjusted R-squared:  0.6658
## F-statistic:  55.8 on 4 and 106 DF,  p-value: < 2.2e-16
```

Conclusion: Interpretation of Final Model

In the final model, I removed median income as a predictor variable, added an interaction term of average income and GDP, and applied a square root transformation on the response and predictor variables.

From the regression summary, we see that the estimated regression function is $\hat{Y} = 0.880053 + 0.018512X_1 + 0.063386X_2 + 0.809363X_3 - 0.012548X_1X_3$.

The interpretation of b_0 is that 0.880053 is the mean value that we would predict for the average satisfaction level if $X_1 = X_2 = X_3 = 0$. It is also known as the y-intercept.

The interpretation of b_1 is that with a one unit increase in X_1 (average income), there will be an increase of 0.018512 to the average satisfaction level while holding X_2 and X_3 constant.

The interpretation of b_2 is that with a one unit increase in X_2 (income inequality), there will be an increase of 0.063386 to the average satisfaction level while holding X_1 and X_3 constant.

The interpretation of b_3 is that with a one unit increase in X_3 (GDP), there will be an increase of 0.809363 to the average satisfaction level while holding X_1 and X_2 constant.

The interpretation of b_1b_3 is that with a one unit increase in either X_1 (average income) or X_3 (GDP), there will be a decrease of 0.1550 to the average satisfaction level while holding X_2 and either X_1 or X_3 constant.

This model suggests that average income, income inequality, and GDP are significant indicators of average satisfaction for countries, with GDP being the strongest indicator. It is also important that we subtract from the average satisfaction score in consideration of the interaction term of GDP and average income. This model shows that income does have an effect on average satisfaction level of countries, and that it is important to consider the many different aspects of income such as average income, income inequality, and GDP.