

# Stat632 Project

Brandon Keck

## World Happiness Index

### A. Introduction

Research goal: We aim to predict a country's Happiness Score using economic and social indicators.

### B. Data Description

```
library(readr)
WHI <- read_csv("WHI_Inflation.csv")
# head(WHI)

WHI <- clean_names(WHI)

# sort(WHI$score, decreasing = TRUE)

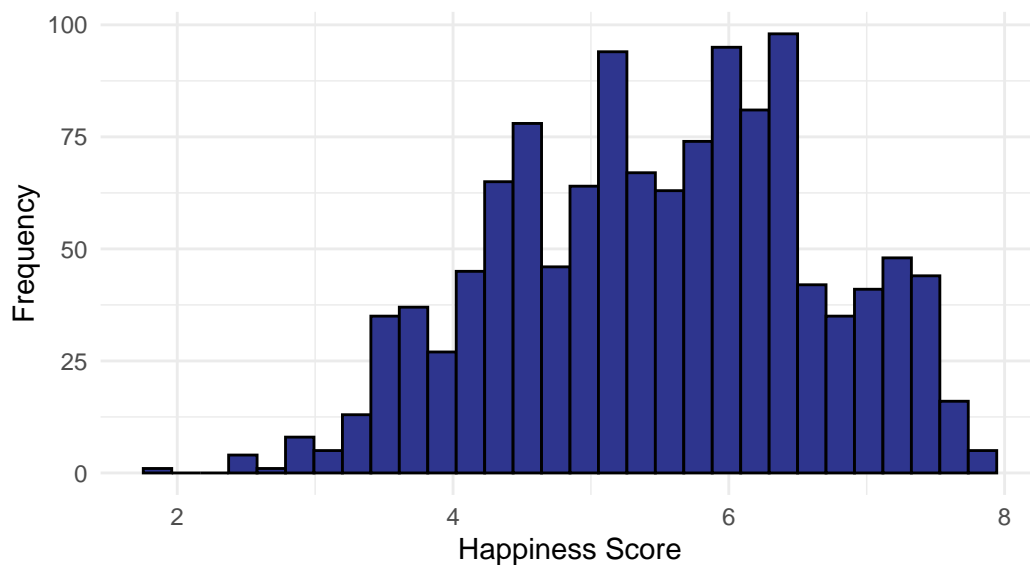
WHI$year <- as.factor(WHI$year)

ggplot(WHI, aes(x = score)) +
  geom_histogram(fill = "#2E358E", color = "black") +
  labs(title = "Figure 1.1",
        subtitle = "Distribution of World Happiness Scores",
        x = "Happiness Score",
        y = "Frequency") +
  theme_minimal()
```

``stat_bin()`` using ``bins = 30``. Pick better value with ``binwidth``.

Figure 1.1

Distribution of World Happiness Scores



```
# ggplot(WHI, aes(x = score, color = year)) +
#   geom_histogram() +
#   facet_wrap(~ year)
```

```
range(WHI$score)
```

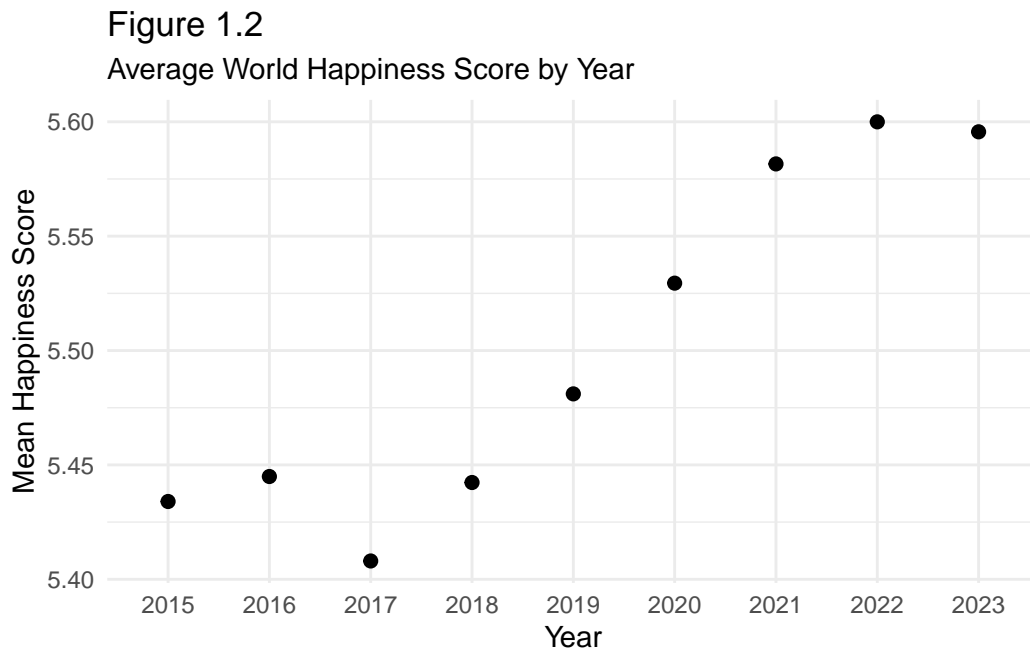
```
[1] 1.859 7.842
```

```
WHI_means <- WHI %>%
  group_by(year) %>%
  summarise(mean(score))
WHI_means
```

```
# A tibble: 9 x 2
  year `mean(score)`
  <fct>      <dbl>
1 2015      5.43
2 2016      5.44
3 2017      5.41
4 2018      5.44
```

5	2019	5.48
6	2020	5.53
7	2021	5.58
8	2022	5.60
9	2023	5.60

```
ggplot(WHI_means, aes(x = year, y = `mean(score)`)) +
  geom_point(size = 2) +
  labs(title = "Figure 1.2",
        subtitle = "Average World Happiness Score by Year",
        x = "Year",
        y = "Mean Happiness Score") +
  theme_minimal()
```



We see a slight left skew in the happiness scores for countries. More countries have higher happiness scores. The mean happiness scores are increasing from 2015 to 2023.

```
head(WHI)
```

```
# A tibble: 6 x 16
  country      year headline_consumer_price_inflation energy_consumer_price_in~1
```

```

      <chr>      <fct>                                <dbl>                                <dbl>
1 Afghanistan 2015                                -0.660                                -4.25
2 Afghanistan 2016                                 4.38                                  2.07
3 Afghanistan 2017                                 4.98                                  4.44
4 Afghanistan 2018                                 0.63                                  1.47
5 Afghanistan 2019                                 2.30                                  -2.49
6 Afghanistan 2020                                 5.44                                  NA
# i abbreviated name: 1: energy_consumer_price_inflation
# i 12 more variables: food_consumer_price_inflation <dbl>,
#   official_core_consumer_price_inflation <dbl>,
#   producer_price_inflation <dbl>, gdp_deflator_index_growth_rate <dbl>,
#   continent_region <chr>, score <dbl>, gdp_per_capita <dbl>,
#   social_support <dbl>, healthy_life_expectancy_at_birth <dbl>,
#   freedom_to_make_life_choices <dbl>, generosity <dbl>, ...

```

```

WHI <- WHI %>%
  rename(overall_infl = headline_consumer_price_inflation,
         energy_infl = energy_consumer_price_inflation,
         food_infl = food_consumer_price_inflation,
         gdp_deflator = gdp_deflator_index_growth_rate,
         gdp = gdp_per_capita,
         life_expectancy = healthy_life_expectancy_at_birth,
         freedom = freedom_to_make_life_choices,
         corruption = perceptions_of_corruption,
         happy_score = score)

```

```
summary(WHI)
```

```

country      year      overall_infl      energy_infl
Length:1232   2015   :142   Min.    : -3.753   Min.    : -23.8800
Class :character 2018   :142   1st Qu.:  1.402   1st Qu.:  0.6076
Mode  :character 2016   :141   Median :  3.476   Median :  2.7692
                2017   :141   Mean    :  7.395   Mean    :  6.4242
                2019   :141   3rd Qu.:  6.876   3rd Qu.:  7.1013
                2020   :138   Max.    :557.210   Max.    :306.4317
                (Other):387   NA's    :32       NA's    :142
 food_infl      official_core_consumer_price_inflation
Min.    : -22.030   Min.    : -28.619
1st Qu.:  1.264   1st Qu.:  1.042
Median :  3.729   Median :  2.246
Mean    :  8.030   Mean    :  3.513

```

3rd Qu.:	9.294	3rd Qu.:	4.627
Max.	:601.020	Max.	: 58.852
NA's	:102	NA's	:498

producer_price_inflation	gdp_deflator	continent_region	happy_score
Min.	:-83.3398	Min.	:-26.100
1st Qu.:	-0.2834	1st Qu.:	1.353
Median :	2.7293	Median :	3.244
Mean :	5.8419	Mean :	7.070
3rd Qu.:	8.4334	3rd Qu.:	7.080
Max.	:128.4766	Max.	:812.247
NA's	:463	NA's	:21

gdp	social_support	life_expectancy	freedom
Min.	:0.000	Min.	:0.0000
1st Qu.:	0.737	1st Qu.:	0.8599
Median :	1.052	Median :	1.0935
Mean :	1.031	Mean :	1.0566
3rd Qu.:	1.343	3rd Qu.:	1.3138
Max.	:2.209	Max.	:1.6440

life_expectancy	freedom
Min.	:0.0000
1st Qu.:	0.4078
Median :	0.6178
Mean :	0.5888
3rd Qu.:	0.7815
Max.	:1.1410

generosity	corruption
Min.	:0.0000
1st Qu.:	0.1170
Median :	0.1830
Mean :	0.1961
3rd Qu.:	0.2520
Max.	:0.8381

corruption	
Min.	:0.0000
1st Qu.:	0.0559
Median :	0.0980
Mean :	0.1335
3rd Qu.:	0.1710
Max.	:0.5870
NA's	:1

```
# Clean the data first
WHI_clean <- na.omit(WHI)

# Then fit models on the cleaned data
lm_full_WHI <- lm(happy_score ~ ., data = WHI_clean)
sum_lm_full_WHI <- summary(lm_full_WHI)

# Stepwise selection
lm_WHI <- step(lm_full_WHI, trace = 0)
sum_lm_WHI <- summary(lm_WHI)

c(sum_lm_full_WHI$r.squared, sum_lm_WHI$r.squared)
```

```
[1] 0.9623298 0.9622836
```

```
c(sum_lm_full_WHI$adj.r.squared, sum_lm_WHI$adj.r.squared)
```

```
[1] 0.9549681 0.9553767
```

```
anova(lm_WHI, lm_full_WHI) # partial F-test
```

#### Analysis of Variance Table

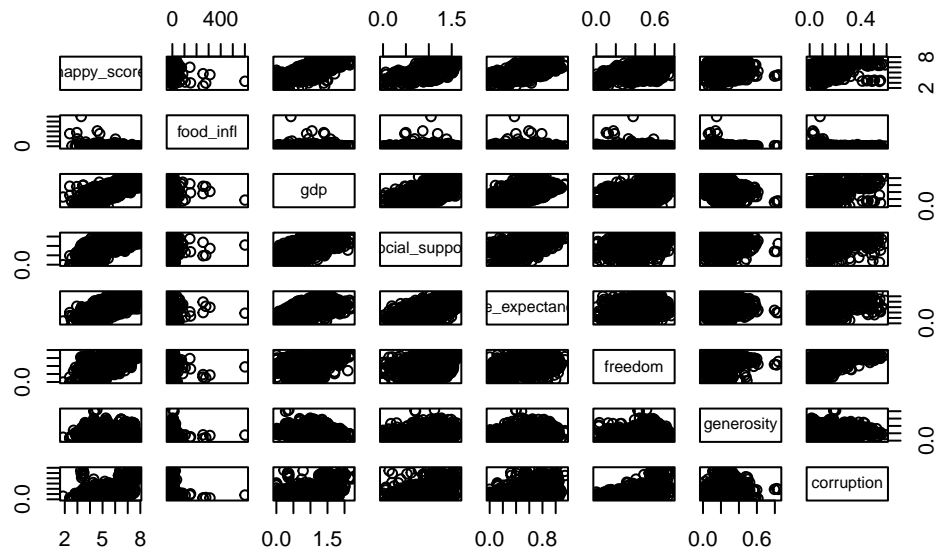
Model 1: happy\_score ~ country + year + official\_core\_consumer\_price\_inflation +  
gdp\_deflator + gdp + social\_support + life\_expectancy + freedom +  
generosity

Model 2: happy\_score ~ country + year + overall\_infl + energy\_infl + food\_infl +  
official\_core\_consumer\_price\_inflation + producer\_price\_inflation +  
gdp\_deflator + continent\_region + gdp + social\_support +  
life\_expectancy + freedom + generosity + corruption

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	486	18.755				
2	481	18.732	5	0.022957	0.1179	0.9884

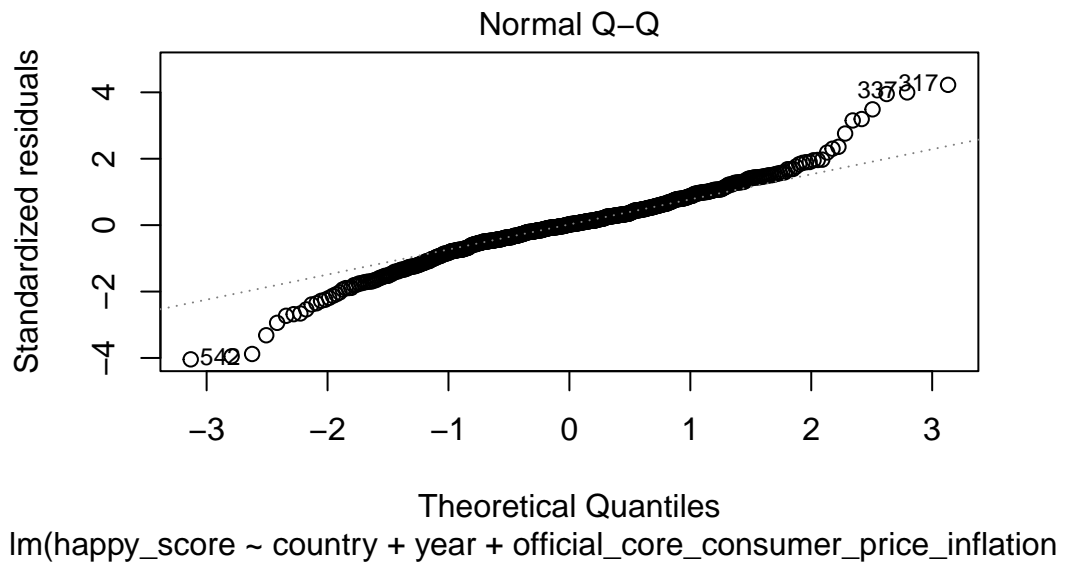
When we remove ‘overall\_infl’, ‘energy\_infl’, and ‘gdp\_deflator’, ‘food\_infl’ becomes more significant. 7 predictors are concluded to be significant. The partial F-test tells us that the reduced model is a sufficient model. We can remove the other predictors.

```
# reduced model
pairs(happy_score ~ food_infl + gdp + social_support + life_expectancy +
      freedom + generosity + corruption, data = WHI)
```



```
plot(lm_WHI, which = 2)
```

Warning: not plotting observations with leverage one:  
389

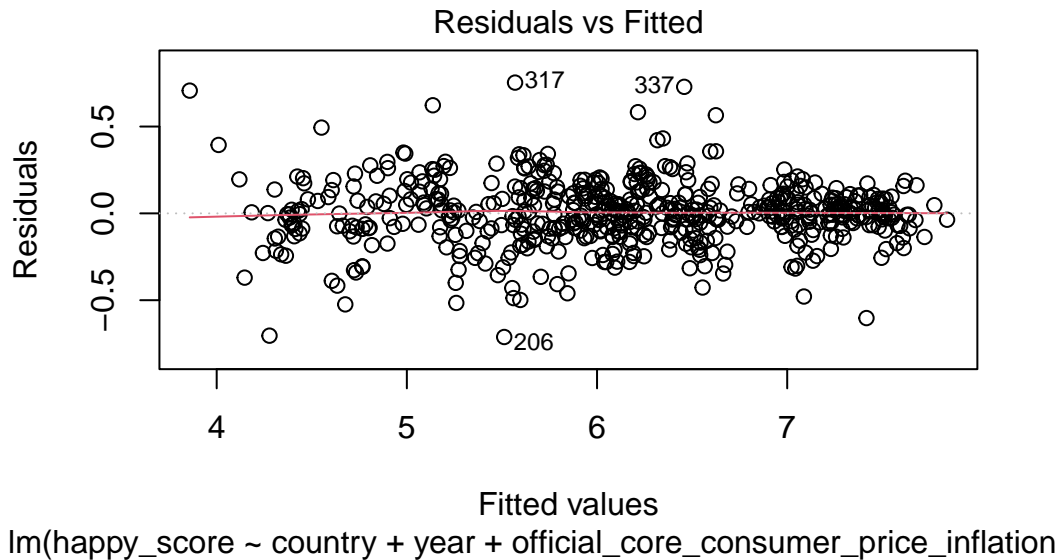


```
shapiro.test(resid(lm_WHI)) # not normal
```

Shapiro-Wilk normality test

```
data:  resid(lm_WHI)
W = 0.96888, p-value = 1.051e-09
```

```
plot(lm_WHI, which = 1)
```



```
library(car)
```

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

```
summary(powerTransform(lm_WHI))
```

bcPower Transformation to Normality

Est Power Rounded Pwr Wald Lwr Bnd Wald Up Bnd

Y1	2.6505	2.65	2.2446	3.0564
----	--------	------	--------	--------

Likelihood ratio test that transformation parameter is equal to 0  
(log transformation)

	LRT	df	pval
LR test, lambda = (0)	190.206	1	< 2.22e-16

Likelihood ratio test that no transformation is needed

	LRT	df	pval
LR test, lambda = (1)	70.33233	1	< 2.22e-16

Start from scratch: Keep all variables but remove missing values. Does anything change?

```
WHI <- read_csv("WHI_Inflation.csv")
WHI <- clean_names(WHI)
WHI <- WHI %>%
  select(-country, -year, -continent_region) %>% # remove ID variables
  na.omit() # remove missing values
# gg_miss_var(WHI) # show number of NA values
# gg_miss_var(WHI, show_pct = TRUE)

WHI <- WHI %>%
  rename(overall_infl = headline_consumer_price_inflation,
         energy_infl = energy_consumer_price_inflation,
         food_infl = food_consumer_price_inflation,
         consumer_infl = official_core_consumer_price_inflation,
         producer_infl = producer_price_inflation,
         gdp_deflator = gdp_deflator_index_growth_rate,
         gdp = gdp_per_capita,
         life_expectancy = healthy_life_expectancy_at_birth,
         freedom = freedom_to_make_life_choices,
         corruption = perceptions_of_corruption,
         happy_score = score)

mean(WHI$happy_score)
```

```
[1] 6.14522
```

```
lm_none_removed <- lm(happy_score ~., data = WHI)
summary(lm_none_removed)
```

Call:

```
lm(formula = happy_score ~ ., data = WHI)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.58949	-0.28855	0.03391	0.30407	1.54843

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	2.397461	0.160369	14.950	< 2e-16	***
overall_infl	-0.020324	0.013041	-1.558	0.1197	
energy_infl	0.006032	0.002660	2.268	0.0237	*
food_infl	-0.009402	0.006989	-1.345	0.1791	
consumer_infl	0.007421	0.007445	0.997	0.3193	
producer_infl	0.002252	0.002738	0.822	0.4112	
gdp_deflator	-0.003818	0.006332	-0.603	0.5467	
gdp	0.835599	0.084542	9.884	< 2e-16	***
social_support	0.681508	0.094088	7.243	1.45e-12	***
life_expectancy	1.075384	0.157937	6.809	2.53e-11	***
freedom	1.497545	0.206418	7.255	1.34e-12	***
generosity	0.861526	0.203714	4.229	2.74e-05	***
corruption	1.442367	0.216147	6.673	6.01e-11	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4975 on 563 degrees of freedom

Multiple R-squared: 0.7198, Adjusted R-squared: 0.7138

F-statistic: 120.5 on 12 and 563 DF, p-value: < 2.2e-16

```
best_none_removed <- step(lm_none_removed, trace = 0)
summary(best_none_removed)
```

Call:

```
lm(formula = happy_score ~ overall_infl + energy_infl + gdp +
    social_support + life_expectancy + freedom + generosity +
    corruption, data = WHI)
```

Residuals:

	Min	1Q	Median	3Q	Max
--	-----	----	--------	----	-----

-1.56256 -0.28652 0.03242 0.29975 1.51134

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.432880	0.155060	15.690	< 2e-16 ***
overall_infl	-0.026145	0.005367	-4.871	1.44e-06 ***
energy_infl	0.006760	0.002495	2.710	0.00693 **
gdp	0.813666	0.082417	9.873	< 2e-16 ***
social_support	0.673631	0.093047	7.240	1.47e-12 ***
life_expectancy	1.068933	0.154902	6.901	1.39e-11 ***
freedom	1.483459	0.203208	7.300	9.77e-13 ***
generosity	0.883452	0.202268	4.368	1.49e-05 ***
corruption	1.460013	0.212579	6.868	1.72e-11 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4972 on 567 degrees of freedom

Multiple R-squared: 0.7181, Adjusted R-squared: 0.7141

F-statistic: 180.6 on 8 and 567 DF, p-value: < 2.2e-16

```
anova(best_none_removed, lm_none_removed) # partial F-test
```

Analysis of Variance Table

Model 1: happy\_score ~ overall\_infl + energy\_infl + gdp + social\_support +  
life\_expectancy + freedom + generosity + corruption

Model 2: happy\_score ~ overall\_infl + energy\_infl + food\_infl + consumer\_infl +  
producer\_infl + gdp\_deflator + gdp + social\_support + life\_expectancy +  
freedom + generosity + corruption

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	567	140.17				
2	563	139.33	4	0.84022	0.8488	0.4946

```
# top 5 happiness scores
```

```
WHI <- read_csv("WHI_Inflation.csv")
```

Rows: 1232 Columns: 16

-- Column specification -----

Delimiter: ","

chr (2): Country, Continent/Region

dbl (14): Year, Headline Consumer Price Inflation, Energy Consumer Price Inf...

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

```
WHI <- clean_names(WHI)
WHI <- WHI %>%
  rename(overall_infl = headline_consumer_price_inflation,
         energy_infl = energy_consumer_price_inflation,
         food_infl = food_consumer_price_inflation,
         consumer_infl = official_core_consumer_price_inflation,
         producer_infl = producer_price_inflation,
         gdp_deflator = gdp_deflator_index_growth_rate,
         gdp = gdp_per_capita,
         life_expectancy = healthy_life_expectancy_at_birth,
         freedom = freedom_to_make_life_choices,
         corruption = perceptions_of_corruption,
         happy_score = score)

score_summary <- WHI %>%
  group_by(country) %>%
  summarise(mean_score = mean(happy_score, na.rm = TRUE)) %>%
  arrange(mean_score)
score_summary
```

# A tibble: 148 x 2

	country	mean_score
	<chr>	<dbl>
1	Afghanistan	2.99
2	Central African Republic	3.20
3	South Sudan	3.27
4	Burundi	3.28
5	Rwanda	3.40
6	Tanzania	3.54
7	Zimbabwe	3.63
8	Botswana	3.67
9	Malawi	3.76
10	Togo	3.81

# i 138 more rows

| Country | Happiness Score |

```
|:-----:|:-----:|
| Finland | 7.663 |
| Denmark | 7.580 |
| Iceland | 7.522 |
| Switzerland | 7.493 |
| Norway | 7.474 |
| Country | Happiness Score |
|:-----:|:-----:|
| Afghanistan | 2.991 |
| Central African Republic | 3.203 |
| South Sudan | 3.269 |
| Burundi | 3.278 |
| Rwanda | 3.399 |
# transform model

lm_trans <- lm((happy_score^2) ~ overall_infl + energy_infl + gdp + social_support +
  life_expectancy + freedom + generosity + corruption, data = WHI)
summary(lm_trans)
```

Call:

```
lm(formula = (happy_score^2) ~ overall_infl + energy_infl + gdp +
  social_support + life_expectancy + freedom + generosity +
  corruption, data = WHI)
```

Residuals:

Min	1Q	Median	3Q	Max
-18.6715	-3.6835	0.2051	4.1004	17.1271

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-5.034005	0.812905	-6.193	8.40e-10 ***
overall_infl	-0.030083	0.013996	-2.149	0.0318 *
energy_infl	-0.007899	0.019486	-0.405	0.6853
gdp	9.833041	0.608908	16.149	< 2e-16 ***

```

social_support    7.334396    0.751908    9.754 < 2e-16 ***
life_expectancy  14.352520    1.012956   14.169 < 2e-16 ***
freedom          12.716007    1.536795    8.274 3.77e-16 ***
generosity       11.157723    1.727606    6.458 1.60e-10 ***
corruption       17.822939    1.870934    9.526 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.094 on 1080 degrees of freedom
(143 observations deleted due to missingness)
Multiple R-squared:  0.7664,    Adjusted R-squared:  0.7647
F-statistic: 442.9 on 8 and 1080 DF,  p-value: < 2.2e-16

```

```
shapiro.test(resid(lm_trans))
```

Shapiro-Wilk normality test

```

data:  resid(lm_trans)
W = 0.99373, p-value = 0.0001538

```

```
# bptest(lm_trans)
```

```
# polynomial model
```

```
library(lmtest)
```

Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

```
as.Date, as.Date.numeric
```

```

WHI_clean <- WHI %>%
  select(-country, -year, -continent_region) %>%

```

```

na.omit()

lm_poly <- lm(happy_score ~ overall_infl + energy_infl + food_infl + consumer_infl + produ

summary(lm_poly)

```

Call:

```

lm(formula = happy_score ~ overall_infl + energy_infl + food_infl +
    consumer_infl + producer_infl + gdp_deflator + gdp + social_support +
    life_expectancy + freedom + generosity + corruption + I(overall_infl^2) +
    I(energy_infl^2) + I(food_infl^2) + I(consumer_infl^2) +
    I(producer_infl^2) + I(gdp_deflator^2) + I(gdp^2) + I(social_support^2) +
    I(life_expectancy^2) + I(freedom^2) + I(generosity^2) + I(corruption^2),
    data = WHI_clean)

```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.80061	-0.28446	0.03086	0.29461	1.36110

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	9.820e-01	4.596e-01	2.137	0.033065	*
overall_infl	-3.930e-02	1.823e-02	-2.156	0.031535	*
energy_infl	1.014e-02	4.457e-03	2.276	0.023252	*
food_infl	-2.405e-03	9.391e-03	-0.256	0.798005	
consumer_infl	9.585e-03	7.524e-03	1.274	0.203246	
producer_infl	1.816e-03	2.780e-03	0.653	0.513811	
gdp_deflator	-4.282e-03	7.681e-03	-0.557	0.577443	
gdp	2.155e+00	4.060e-01	5.308	1.61e-07	***
social_support	7.877e-01	5.073e-01	1.553	0.121063	
life_expectancy	2.484e+00	7.910e-01	3.140	0.001779	**
freedom	1.827e+00	9.035e-01	2.023	0.043601	*
generosity	1.162e+00	6.345e-01	1.831	0.067599	.
corruption	3.866e+00	5.774e-01	6.696	5.30e-11	***
I(overall_infl^2)	3.318e-04	5.779e-04	0.574	0.566098	
I(energy_infl^2)	-3.731e-05	8.392e-05	-0.445	0.656765	
I(food_infl^2)	-2.205e-04	3.035e-04	-0.726	0.467904	
I(consumer_infl^2)	3.061e-04	3.632e-04	0.843	0.399814	
I(producer_infl^2)	-4.954e-05	5.218e-05	-0.949	0.342824	
I(gdp_deflator^2)	5.342e-05	2.081e-04	0.257	0.797506	
I(gdp^2)	-5.114e-01	1.583e-01	-3.231	0.001305	**

```

I(social_support^2) -3.607e-02  2.289e-01  -0.158  0.874829
I(life_expectancy^2) -1.250e+00  5.500e-01  -2.272  0.023455 *
I(freedom^2)        -4.551e-01  1.017e+00  -0.447  0.654721
I(generosity^2)     -8.908e-01  1.206e+00  -0.739  0.460443
I(corruption^2)     -4.554e+00  1.186e+00  -3.840  0.000137 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4834 on 551 degrees of freedom
Multiple R-squared:  0.7411,    Adjusted R-squared:  0.7298
F-statistic: 65.72 on 24 and 551 DF,  p-value: < 2.2e-16

```

```
shapiro.test(resid(lm_poly))
```

Shapiro-Wilk normality test

```

data:  resid(lm_poly)
W = 0.9882, p-value = 0.0001339

```

```
bptest(lm_poly) # alpha = 0.01
```

studentized Breusch-Pagan test

```

data:  lm_poly
BP = 41.309, df = 24, p-value = 0.0154

```

```

best_poly <- step(lm_poly, trace = 0)
summary(best_poly)

```

Call:

```

lm(formula = happy_score ~ overall_infl + energy_infl + consumer_infl +
    gdp + social_support + life_expectancy + freedom + generosity +
    corruption + I(food_infl^2) + I(consumer_infl^2) + I(gdp^2) +
    I(life_expectancy^2) + I(corruption^2), data = WHI_clean)

```

Residuals:

Min	1Q	Median	3Q	Max
-1.78560	-0.27610	0.03327	0.29574	1.35947

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	1.0952358	0.3367488	3.252	0.001213	**
overall_infl	-0.0393053	0.0104496	-3.761	0.000187	***
energy_infl	0.0090691	0.0025260	3.590	0.000359	***
consumer_infl	0.0107462	0.0072942	1.473	0.141242	
gdp	2.1740531	0.3912891	5.556	4.27e-08	***
social_support	0.7049974	0.0917926	7.680	7.11e-14	***
life_expectancy	2.4570808	0.7570702	3.246	0.001242	**
freedom	1.4742908	0.1987181	7.419	4.39e-13	***
generosity	0.7405350	0.1980755	3.739	0.000204	***
corruption	3.9589778	0.5574034	7.103	3.73e-12	***
I(food_infl^2)	-0.0002281	0.0001440	-1.585	0.113602	
I(consumer_infl^2)	0.0004867	0.0002842	1.713	0.087327	.
I(gdp^2)	-0.5324256	0.1516904	-3.510	0.000484	***
I(life_expectancy^2)	-1.2015135	0.5264687	-2.282	0.022850	*
I(corruption^2)	-4.7880285	1.1130824	-4.302	2.00e-05	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4804 on 561 degrees of freedom

Multiple R-squared: 0.7396, Adjusted R-squared: 0.7331

F-statistic: 113.8 on 14 and 561 DF, p-value: < 2.2e-16

```
shapiro.test(resid(best_poly))
```

Shapiro-Wilk normality test

data: resid(best\_poly)

W = 0.98886, p-value = 0.0002293

```
bptest(best_poly)
```

studentized Breusch-Pagan test

```
data: best_poly
BP = 38.217, df = 14, p-value = 0.0004813
```

```
final_lm <- lm(happy_score ~ overall_infl + energy_infl +
  gdp + social_support + life_expectancy + freedom + generosity +
  corruption + I(gdp^2) +
  I(life_expectancy^2) + I(corruption^2), data = WHI)
summary(final_lm)
```

Call:

```
lm(formula = happy_score ~ overall_infl + energy_infl + gdp +
  social_support + life_expectancy + freedom + generosity +
  corruption + I(gdp^2) + I(life_expectancy^2) + I(corruption^2),
  data = WHI)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.88563	-0.31963	0.01196	0.35116	1.48254

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.324720	0.117096	19.853	< 2e-16 ***
overall_infl	-0.002792	0.001268	-2.201	0.027934 *
energy_infl	-0.002261	0.001774	-1.275	0.202620
gdp	0.862348	0.177358	4.862	1.33e-06 ***
social_support	0.636321	0.068905	9.235	< 2e-16 ***
life_expectancy	0.400198	0.298727	1.340	0.180633
freedom	1.253188	0.139491	8.984	< 2e-16 ***
generosity	0.833943	0.155652	5.358	1.03e-07 ***
corruption	2.704031	0.475636	5.685	1.68e-08 ***
I(gdp^2)	0.056546	0.080629	0.701	0.483259
I(life_expectancy^2)	0.889780	0.258415	3.443	0.000597 ***
I(corruption^2)	-4.147410	0.982232	-4.222	2.62e-05 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5483 on 1077 degrees of freedom

(143 observations deleted due to missingness)

Multiple R-squared: 0.7697, Adjusted R-squared: 0.7673

F-statistic: 327.2 on 11 and 1077 DF, p-value: < 2.2e-16

## Trying a Log Transformation Instead

```
lm_trans2 <- lm(log(happy_score) ~ overall_infl + energy_infl + gdp + social_support +  
  life_expectancy + freedom + generosity + corruption, data = WHI)  
summary(lm_trans2)
```

Call:

```
lm(formula = log(happy_score) ~ overall_infl + energy_infl +  
  gdp + social_support + life_expectancy + freedom + generosity +  
  corruption, data = WHI)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.41133	-0.05855	0.00721	0.06801	0.28759

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.0654611	0.0145420	73.268	< 2e-16 ***
overall_infl	-0.0006755	0.0002504	-2.698	0.00709 **
energy_infl	-0.0006875	0.0003486	-1.972	0.04884 *
gdp	0.1824827	0.0108927	16.753	< 2e-16 ***
social_support	0.1289043	0.0134508	9.583	< 2e-16 ***
life_expectancy	0.2594724	0.0181207	14.319	< 2e-16 ***
freedom	0.2432725	0.0274916	8.849	< 2e-16 ***
generosity	0.1310289	0.0309050	4.240	2.43e-05 ***
corruption	0.0740459	0.0334690	2.212	0.02715 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.109 on 1080 degrees of freedom

(143 observations deleted due to missingness)

Multiple R-squared: 0.7469, Adjusted R-squared: 0.745

F-statistic: 398.3 on 8 and 1080 DF, p-value: < 2.2e-16

```
shapiro.test(resid(lm_trans2))
```

Shapiro-Wilk normality test

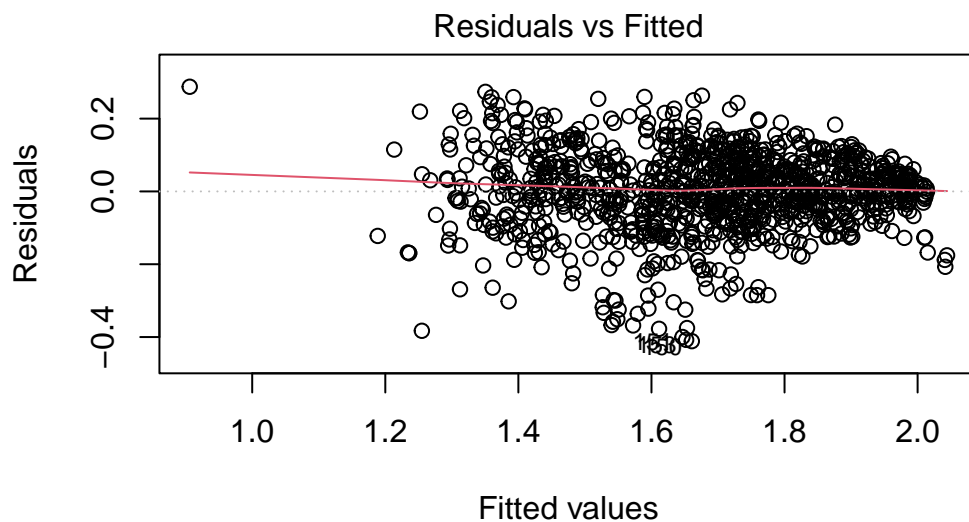
```
data: resid(lm_trans2)
W = 0.9752, p-value = 1.055e-12
```

```
bptest(lm_trans2)
```

studentized Breusch-Pagan test

```
data: lm_trans2
BP = 130.04, df = 8, p-value < 2.2e-16
```

```
plot(lm_trans2, 1)
```



$\text{lm}(\log(\text{happy\_score}) \sim \text{overall\_infl} + \text{energy\_infl} + \text{gdp} + \text{social\_support} + \dots)$

**Let's do some Poly**

```
lm_cubic <- lm(happy_score ~
  poly(overall_infl, 3, raw = TRUE) +
  poly(energy_infl, 3, raw = TRUE) +
  poly(gdp, 3, raw = TRUE) +
```

```

poly(social_support, 3, raw = TRUE) +
poly(life_expectancy, 3, raw = TRUE) +
poly(freedom, 3, raw = TRUE) +
poly(generosity, 3, raw = TRUE) +
poly(corruption, 3, raw = TRUE),
data = WHI_clean
)

summary(lm_cubic)

```

Call:

```

lm(formula = happy_score ~ poly(overall_infl, 3, raw = TRUE) +
    poly(energy_infl, 3, raw = TRUE) + poly(gdp, 3, raw = TRUE) +
    poly(social_support, 3, raw = TRUE) + poly(life_expectancy,
    3, raw = TRUE) + poly(freedom, 3, raw = TRUE) + poly(generosity,
    3, raw = TRUE) + poly(corruption, 3, raw = TRUE), data = WHI_clean)

```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.80343	-0.28822	0.04019	0.28980	1.34520

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.696e+00	9.830e-01	1.726	0.08499 .
poly(overall_infl, 3, raw = TRUE)1	-7.491e-03	1.771e-02	-0.423	0.67239
poly(overall_infl, 3, raw = TRUE)2	-2.245e-03	1.293e-03	-1.736	0.08304 .
poly(overall_infl, 3, raw = TRUE)3	4.093e-05	2.086e-05	1.963	0.05019 .
poly(energy_infl, 3, raw = TRUE)1	6.900e-03	4.594e-03	1.502	0.13372
poly(energy_infl, 3, raw = TRUE)2	2.366e-04	2.012e-04	1.176	0.23993
poly(energy_infl, 3, raw = TRUE)3	-4.363e-06	2.443e-06	-1.786	0.07468 .
poly(gdp, 3, raw = TRUE)1	-5.037e-01	1.310e+00	-0.384	0.70082
poly(gdp, 3, raw = TRUE)2	1.773e+00	1.065e+00	1.664	0.09661 .
poly(gdp, 3, raw = TRUE)3	-6.154e-01	2.762e-01	-2.228	0.02629 *
poly(social_support, 3, raw = TRUE)1	7.587e-01	1.761e+00	0.431	0.66677
poly(social_support, 3, raw = TRUE)2	2.218e-03	1.822e+00	0.001	0.99903
poly(social_support, 3, raw = TRUE)3	-1.828e-02	5.988e-01	-0.031	0.97565
poly(life_expectancy, 3, raw = TRUE)1	1.958e-01	2.756e+00	0.071	0.94341
poly(life_expectancy, 3, raw = TRUE)2	2.544e+00	4.271e+00	0.596	0.55159
poly(life_expectancy, 3, raw = TRUE)3	-1.868e+00	2.100e+00	-0.890	0.37411
poly(freedom, 3, raw = TRUE)1	7.721e+00	2.798e+00	2.759	0.00599 **

```

poly(freedom, 3, raw = TRUE)2      -1.588e+01  6.899e+00  -2.302  0.02173 *
poly(freedom, 3, raw = TRUE)3      1.218e+01  5.360e+00   2.272  0.02348 *
poly(generosity, 3, raw = TRUE)1    4.811e-01  1.460e+00   0.329  0.74194
poly(generosity, 3, raw = TRUE)2    1.803e+00  6.296e+00   0.286  0.77470
poly(generosity, 3, raw = TRUE)3   -2.839e+00  7.709e+00  -0.368  0.71283
poly(corruption, 3, raw = TRUE)1    3.740e+00  1.162e+00   3.217  0.00137 **
poly(corruption, 3, raw = TRUE)2   -4.170e+00  5.589e+00  -0.746  0.45590
poly(corruption, 3, raw = TRUE)3   -6.038e-01  7.428e+00  -0.081  0.93525

```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4793 on 551 degrees of freedom

Multiple R-squared: 0.7454, Adjusted R-squared: 0.7343

F-statistic: 67.22 on 24 and 551 DF, p-value: < 2.2e-16

```
shapiro.test(resid(lm_cubic))
```

Shapiro-Wilk normality test

data: resid(lm\_cubic)

W = 0.99015, p-value = 0.0006712

```
bptest(lm_cubic) # alpha = 0.01
```

studentized Breusch-Pagan test

data: lm\_cubic

BP = 57.997, df = 24, p-value = 0.0001218

```
best_cubic <- step(lm_cubic, trace = 0)
summary(best_cubic)
```

Call:

```
lm(formula = happy_score ~ poly(overall_infl, 3, raw = TRUE) +
    poly(energy_infl, 3, raw = TRUE) + poly(gdp, 3, raw = TRUE) +
```

```
poly(social_support, 3, raw = TRUE) + poly(life_expectancy,
3, raw = TRUE) + poly(freedom, 3, raw = TRUE) + poly(generosity,
3, raw = TRUE) + poly(corruption, 3, raw = TRUE), data = WHI_clean)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.80343	-0.28822	0.04019	0.28980	1.34520

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.696e+00	9.830e-01	1.726	0.08499 .
poly(overall_infl, 3, raw = TRUE)1	-7.491e-03	1.771e-02	-0.423	0.67239
poly(overall_infl, 3, raw = TRUE)2	-2.245e-03	1.293e-03	-1.736	0.08304 .
poly(overall_infl, 3, raw = TRUE)3	4.093e-05	2.086e-05	1.963	0.05019 .
poly(energy_infl, 3, raw = TRUE)1	6.900e-03	4.594e-03	1.502	0.13372
poly(energy_infl, 3, raw = TRUE)2	2.366e-04	2.012e-04	1.176	0.23993
poly(energy_infl, 3, raw = TRUE)3	-4.363e-06	2.443e-06	-1.786	0.07468 .
poly(gdp, 3, raw = TRUE)1	-5.037e-01	1.310e+00	-0.384	0.70082
poly(gdp, 3, raw = TRUE)2	1.773e+00	1.065e+00	1.664	0.09661 .
poly(gdp, 3, raw = TRUE)3	-6.154e-01	2.762e-01	-2.228	0.02629 *
poly(social_support, 3, raw = TRUE)1	7.587e-01	1.761e+00	0.431	0.66677
poly(social_support, 3, raw = TRUE)2	2.218e-03	1.822e+00	0.001	0.99903
poly(social_support, 3, raw = TRUE)3	-1.828e-02	5.988e-01	-0.031	0.97565
poly(life_expectancy, 3, raw = TRUE)1	1.958e-01	2.756e+00	0.071	0.94341
poly(life_expectancy, 3, raw = TRUE)2	2.544e+00	4.271e+00	0.596	0.55159
poly(life_expectancy, 3, raw = TRUE)3	-1.868e+00	2.100e+00	-0.890	0.37411
poly(freedom, 3, raw = TRUE)1	7.721e+00	2.798e+00	2.759	0.00599 **
poly(freedom, 3, raw = TRUE)2	-1.588e+01	6.899e+00	-2.302	0.02173 *
poly(freedom, 3, raw = TRUE)3	1.218e+01	5.360e+00	2.272	0.02348 *
poly(generosity, 3, raw = TRUE)1	4.811e-01	1.460e+00	0.329	0.74194
poly(generosity, 3, raw = TRUE)2	1.803e+00	6.296e+00	0.286	0.77470
poly(generosity, 3, raw = TRUE)3	-2.839e+00	7.709e+00	-0.368	0.71283
poly(corruption, 3, raw = TRUE)1	3.740e+00	1.162e+00	3.217	0.00137 **
poly(corruption, 3, raw = TRUE)2	-4.170e+00	5.589e+00	-0.746	0.45590
poly(corruption, 3, raw = TRUE)3	-6.038e-01	7.428e+00	-0.081	0.93525

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4793 on 551 degrees of freedom

Multiple R-squared: 0.7454, Adjusted R-squared: 0.7343

F-statistic: 67.22 on 24 and 551 DF, p-value: < 2.2e-16

```
shapiro.test(resid(best_cubic))
```

Shapiro-Wilk normality test

```
data: resid(best_cubic)
W = 0.99015, p-value = 0.0006712
```

```
bptest(best_cubic) # alpha = 0.01
```

studentized Breusch-Pagan test

```
data: best_cubic
BP = 57.997, df = 24, p-value = 0.0001218
```

## Quartic

```
lm_quartic <- lm(happy_score ~
  poly(overall_infl, 4, raw = TRUE) +
  poly(energy_infl, 4, raw = TRUE) +
  poly(gdp, 4, raw = TRUE) +
  poly(social_support, 4, raw = TRUE) +
  poly(life_expectancy, 4, raw = TRUE) +
  poly(freedom, 4, raw = TRUE) +
  poly(generosity, 4, raw = TRUE) +
  poly(corruption, 4, raw = TRUE),
  data = WHI_clean
)
```

```
summary(lm_quartic)
```

Call:

```
lm(formula = happy_score ~ poly(overall_infl, 4, raw = TRUE) +
  poly(energy_infl, 4, raw = TRUE) + poly(gdp, 4, raw = TRUE) +
  poly(social_support, 4, raw = TRUE) + poly(life_expectancy,
```

```
4, raw = TRUE) + poly(freedom, 4, raw = TRUE) + poly(generosity,
4, raw = TRUE) + poly(corruption, 4, raw = TRUE), data = WHI_clean)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.77813	-0.27722	0.02643	0.28483	1.24079

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.200e+00	1.995e+00	1.103	0.27053
poly(overall_infl, 4, raw = TRUE)1	3.648e-03	2.309e-02	0.158	0.87451
poly(overall_infl, 4, raw = TRUE)2	-2.807e-03	2.715e-03	-1.034	0.30171
poly(overall_infl, 4, raw = TRUE)3	4.878e-05	1.183e-04	0.412	0.68026
poly(overall_infl, 4, raw = TRUE)4	-1.659e-09	1.560e-06	-0.001	0.99915
poly(energy_infl, 4, raw = TRUE)1	5.057e-03	5.047e-03	1.002	0.31680
poly(energy_infl, 4, raw = TRUE)2	2.962e-04	2.538e-04	1.167	0.24379
poly(energy_infl, 4, raw = TRUE)3	-4.887e-06	9.147e-06	-0.534	0.59339
poly(energy_infl, 4, raw = TRUE)4	-1.038e-09	9.615e-08	-0.011	0.99139
poly(gdp, 4, raw = TRUE)1	-6.027e+00	3.430e+00	-1.757	0.07943
poly(gdp, 4, raw = TRUE)2	9.720e+00	4.506e+00	2.157	0.03143
poly(gdp, 4, raw = TRUE)3	-5.225e+00	2.489e+00	-2.099	0.03629
poly(gdp, 4, raw = TRUE)4	9.288e-01	4.894e-01	1.898	0.05826
poly(social_support, 4, raw = TRUE)1	-4.156e+00	5.429e+00	-0.765	0.44433
poly(social_support, 4, raw = TRUE)2	9.185e+00	9.407e+00	0.976	0.32933
poly(social_support, 4, raw = TRUE)3	-6.858e+00	6.748e+00	-1.016	0.30998
poly(social_support, 4, raw = TRUE)4	1.768e+00	1.717e+00	1.030	0.30365
poly(life_expectancy, 4, raw = TRUE)1	1.800e+01	8.994e+00	2.002	0.04582
poly(life_expectancy, 4, raw = TRUE)2	-4.231e+01	2.251e+01	-1.879	0.06072
poly(life_expectancy, 4, raw = TRUE)3	4.520e+01	2.348e+01	1.925	0.05478
poly(life_expectancy, 4, raw = TRUE)4	-1.755e+01	8.724e+00	-2.012	0.04471
poly(freedom, 4, raw = TRUE)1	-5.747e+00	6.449e+00	-0.891	0.37320
poly(freedom, 4, raw = TRUE)2	4.316e+01	2.617e+01	1.649	0.09963
poly(freedom, 4, raw = TRUE)3	-8.936e+01	4.401e+01	-2.030	0.04280
poly(freedom, 4, raw = TRUE)4	5.991e+01	2.616e+01	2.291	0.02237
poly(generosity, 4, raw = TRUE)1	-9.567e-01	2.733e+00	-0.350	0.72641
poly(generosity, 4, raw = TRUE)2	1.402e+01	1.957e+01	0.716	0.47399
poly(generosity, 4, raw = TRUE)3	-4.036e+01	5.334e+01	-0.757	0.44962
poly(generosity, 4, raw = TRUE)4	3.652e+01	4.808e+01	0.760	0.44780
poly(corruption, 4, raw = TRUE)1	8.357e+00	2.130e+00	3.923	9.86e-05
poly(corruption, 4, raw = TRUE)2	-4.704e+01	1.767e+01	-2.662	0.00799
poly(corruption, 4, raw = TRUE)3	1.326e+02	5.192e+01	2.554	0.01093
poly(corruption, 4, raw = TRUE)4	-1.294e+02	4.910e+01	-2.636	0.00863

```

(Intercept)
poly(overall_infl, 4, raw = TRUE)1
poly(overall_infl, 4, raw = TRUE)2
poly(overall_infl, 4, raw = TRUE)3
poly(overall_infl, 4, raw = TRUE)4
poly(energy_infl, 4, raw = TRUE)1
poly(energy_infl, 4, raw = TRUE)2
poly(energy_infl, 4, raw = TRUE)3
poly(energy_infl, 4, raw = TRUE)4
poly(gdp, 4, raw = TRUE)1      .
poly(gdp, 4, raw = TRUE)2      *
poly(gdp, 4, raw = TRUE)3      *
poly(gdp, 4, raw = TRUE)4      .
poly(social_support, 4, raw = TRUE)1
poly(social_support, 4, raw = TRUE)2
poly(social_support, 4, raw = TRUE)3
poly(social_support, 4, raw = TRUE)4
poly(life_expectancy, 4, raw = TRUE)1 *
poly(life_expectancy, 4, raw = TRUE)2 .
poly(life_expectancy, 4, raw = TRUE)3 .
poly(life_expectancy, 4, raw = TRUE)4 *
poly(freedom, 4, raw = TRUE)1
poly(freedom, 4, raw = TRUE)2      .
poly(freedom, 4, raw = TRUE)3      *
poly(freedom, 4, raw = TRUE)4      *
poly(generosity, 4, raw = TRUE)1
poly(generosity, 4, raw = TRUE)2
poly(generosity, 4, raw = TRUE)3
poly(generosity, 4, raw = TRUE)4
poly(corruption, 4, raw = TRUE)1    ***
poly(corruption, 4, raw = TRUE)2    **
poly(corruption, 4, raw = TRUE)3    *
poly(corruption, 4, raw = TRUE)4    **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 0.4731 on 543 degrees of freedom
Multiple R-squared:  0.7556,    Adjusted R-squared:  0.7412
F-statistic: 52.47 on 32 and 543 DF,  p-value: < 2.2e-16

```

```
shapiro.test(resid(lm_quartic))
```

Shapiro-Wilk normality test

```
data: resid(lm_quartic)
W = 0.98825, p-value = 0.00014
```

```
bptest(lm_quartic) # alpha = 0.01
```

studentized Breusch-Pagan test

```
data: lm_quartic
BP = 66.818, df = 32, p-value = 0.0002984
```

```
best_quartic <- step(lm_quartic, trace = 0)
summary(best_quartic)
```

Call:

```
lm(formula = happy_score ~ poly(overall_infl, 4, raw = TRUE) +
    poly(gdp, 4, raw = TRUE) + poly(social_support, 4, raw = TRUE) +
    poly(life_expectancy, 4, raw = TRUE) + poly(freedom, 4, raw = TRUE) +
    poly(generosity, 4, raw = TRUE) + poly(corruption, 4, raw = TRUE),
    data = WHI_clean)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.78974	-0.26871	0.02992	0.29280	1.21614

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.606e+00	1.984e+00	1.314	0.18945
poly(overall_infl, 4, raw = TRUE)1	2.150e-02	2.030e-02	1.059	0.28996
poly(overall_infl, 4, raw = TRUE)2	-3.774e-03	2.114e-03	-1.785	0.07480
poly(overall_infl, 4, raw = TRUE)3	9.309e-05	6.044e-05	1.540	0.12410
poly(overall_infl, 4, raw = TRUE)4	-6.561e-07	4.922e-07	-1.333	0.18307
poly(gdp, 4, raw = TRUE)1	-5.786e+00	3.396e+00	-1.704	0.08897
poly(gdp, 4, raw = TRUE)2	9.110e+00	4.453e+00	2.046	0.04126
poly(gdp, 4, raw = TRUE)3	-4.770e+00	2.458e+00	-1.940	0.05287
poly(gdp, 4, raw = TRUE)4	8.297e-01	4.839e-01	1.715	0.08699

poly(social_support, 4, raw = TRUE)1	-4.754e+00	5.418e+00	-0.877	0.38066
poly(social_support, 4, raw = TRUE)2	1.037e+01	9.385e+00	1.105	0.26970
poly(social_support, 4, raw = TRUE)3	-7.785e+00	6.731e+00	-1.157	0.24792
poly(social_support, 4, raw = TRUE)4	2.009e+00	1.713e+00	1.173	0.24142
poly(life_expectancy, 4, raw = TRUE)1	1.550e+01	8.935e+00	1.734	0.08343
poly(life_expectancy, 4, raw = TRUE)2	-3.680e+01	2.240e+01	-1.643	0.10097
poly(life_expectancy, 4, raw = TRUE)3	4.062e+01	2.341e+01	1.735	0.08332
poly(life_expectancy, 4, raw = TRUE)4	-1.627e+01	8.712e+00	-1.868	0.06236
poly(freedom, 4, raw = TRUE)1	-5.733e+00	6.466e+00	-0.887	0.37568
poly(freedom, 4, raw = TRUE)2	4.394e+01	2.624e+01	1.675	0.09458
poly(freedom, 4, raw = TRUE)3	-9.179e+01	4.413e+01	-2.080	0.03798
poly(freedom, 4, raw = TRUE)4	6.186e+01	2.622e+01	2.359	0.01865
poly(generosity, 4, raw = TRUE)1	-9.847e-01	2.737e+00	-0.360	0.71914
poly(generosity, 4, raw = TRUE)2	1.413e+01	1.961e+01	0.721	0.47141
poly(generosity, 4, raw = TRUE)3	-4.112e+01	5.345e+01	-0.769	0.44201
poly(generosity, 4, raw = TRUE)4	3.770e+01	4.818e+01	0.783	0.43419
poly(corruption, 4, raw = TRUE)1	8.597e+00	2.133e+00	4.031	6.33e-05
poly(corruption, 4, raw = TRUE)2	-5.059e+01	1.764e+01	-2.868	0.00429
poly(corruption, 4, raw = TRUE)3	1.457e+02	5.171e+01	2.818	0.00501
poly(corruption, 4, raw = TRUE)4	-1.436e+02	4.878e+01	-2.943	0.00339

(Intercept)

poly(overall_infl, 4, raw = TRUE)1	
poly(overall_infl, 4, raw = TRUE)2	.
poly(overall_infl, 4, raw = TRUE)3	
poly(overall_infl, 4, raw = TRUE)4	
poly(gdp, 4, raw = TRUE)1	.
poly(gdp, 4, raw = TRUE)2	*
poly(gdp, 4, raw = TRUE)3	.
poly(gdp, 4, raw = TRUE)4	.
poly(social_support, 4, raw = TRUE)1	
poly(social_support, 4, raw = TRUE)2	
poly(social_support, 4, raw = TRUE)3	
poly(social_support, 4, raw = TRUE)4	
poly(life_expectancy, 4, raw = TRUE)1	.
poly(life_expectancy, 4, raw = TRUE)2	
poly(life_expectancy, 4, raw = TRUE)3	.
poly(life_expectancy, 4, raw = TRUE)4	.
poly(freedom, 4, raw = TRUE)1	
poly(freedom, 4, raw = TRUE)2	.
poly(freedom, 4, raw = TRUE)3	*
poly(freedom, 4, raw = TRUE)4	*
poly(generosity, 4, raw = TRUE)1	

```

poly(generosity, 4, raw = TRUE)2
poly(generosity, 4, raw = TRUE)3
poly(generosity, 4, raw = TRUE)4
poly(corruption, 4, raw = TRUE)1      ***
poly(corruption, 4, raw = TRUE)2      **
poly(corruption, 4, raw = TRUE)3      **
poly(corruption, 4, raw = TRUE)4      **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4745 on 547 degrees of freedom
Multiple R-squared:  0.7524,    Adjusted R-squared:  0.7397
F-statistic: 59.36 on 28 and 547 DF,  p-value: < 2.2e-16

```

```
shapiro.test(resid(best_quartic))
```

Shapiro-Wilk normality test

```

data:  resid(best_quartic)
W = 0.98728, p-value = 6.531e-05

```

```
bptest(best_quartic) # alpha = 0.01
```

studentized Breusch-Pagan test

```

data:  best_quartic
BP = 67.178, df = 28, p-value = 4.569e-05

```

## Cross Validation

```
set.seed(632) # set seed for reproducibility
```

```
n <- nrow(WHI_clean);n
```

```
[1] 576
```

```
floor(0.7*n)
```

```
[1] 403
```

```
# Randomly sample 70% of rows for training set
train <- sample(1:n, 403)

lm_train <- lm(happy_score ~ overall_infl + energy_infl + gdp + social_support +
               life_expectancy + freedom + generosity + corruption, data = WHI_clean,

summary(lm_train)
```

Call:

```
lm(formula = happy_score ~ overall_infl + energy_infl + gdp +
    social_support + life_expectancy + freedom + generosity +
    corruption, data = WHI_clean, subset = train)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.58388	-0.26945	0.05472	0.29805	1.51184

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	2.507563	0.187681	13.361	< 2e-16	***
overall_infl	-0.029838	0.007898	-3.778	0.000183	***
energy_infl	0.006879	0.003136	2.194	0.028838	*
gdp	0.778147	0.097374	7.991	1.49e-14	***
social_support	0.631846	0.115631	5.464	8.25e-08	***
life_expectancy	1.177855	0.185855	6.338	6.39e-10	***
freedom	1.509492	0.238038	6.341	6.25e-10	***
generosity	0.731287	0.254410	2.874	0.004267	**
corruption	1.353289	0.249811	5.417	1.06e-07	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4951 on 394 degrees of freedom

Multiple R-squared: 0.709, Adjusted R-squared: 0.7031

F-statistic: 120 on 8 and 394 DF, p-value: < 2.2e-16

## Make Predictions on the Test Set

```
# subset data frame for testing observations
WHI_test <- WHI_clean[-train, ]

# make predictions for probabilities on test set
probs_test <- predict(lm_train, newdata = WHI_test)

length(probs_test)
```

```
[1] 173
```

```
preds_test <- rep(0, 173)
preds_test[probs_test > 0.5] <- 1

head(probs_test)
```

```
      1      2      3      4      5      6
5.065445 5.517201 5.552764 5.241232 5.022705 4.929676
```

## Calculate the test performance metrics

```
library(caret) # for cross-validation methods
```

Loading required package: lattice

Attaching package: 'caret'

The following object is masked from 'package:purrr':

```
lift
```

```
predictions <- lm_train %>% predict(WHI_test)
data.frame(R2 = R2(predictions, WHI_test$happy_score),
           RMSE = RMSE(predictions, WHI_test$happy_score),
```

```
MAE = MAE(predictions, WHI_test$happy_score))
```

	R2	RMSE	MAE
1	0.7316706	0.5077402	0.4006595

**Test  $R^2$ :** The model explains about 73.17% of the variation in happiness score on unseen data.

**Test RMSE:** On average our prediction is 0.51 happiness points off from the true value

**Test MAE:** The mean absolute error is about 0.40 points. Solid and consistent with RMSE

After performing a 70/30 cross-validation the final linear model achieved a Test  $R^2$  of 0.7317 indicating that about 73% of the variance in happy\_score can be explained in the new data. The test RMSE was approximately 0.508 meaning predictions are off by about half a happiness point on average. These results suggest the model generalizes well beyond the training data.

## Random Forest

```
WHI_train <- WHI_clean[train, ]  
library(randomForest)
```

```
randomForest 4.7-1.1
```

Type `rfNews()` to see new features/changes/bug fixes.

Attaching package: 'randomForest'

The following object is masked from 'package:dplyr':

```
combine
```

The following object is masked from 'package:ggplot2':

```
margin
```

```

set.seed(632) # set seed for reproducibility
rf1 <- randomForest(happy_score ~ ., data = WHI_train, importance = TRUE)

# predict happiness score on test data
rf_preds <- predict(rf1, newdata = WHI_test)

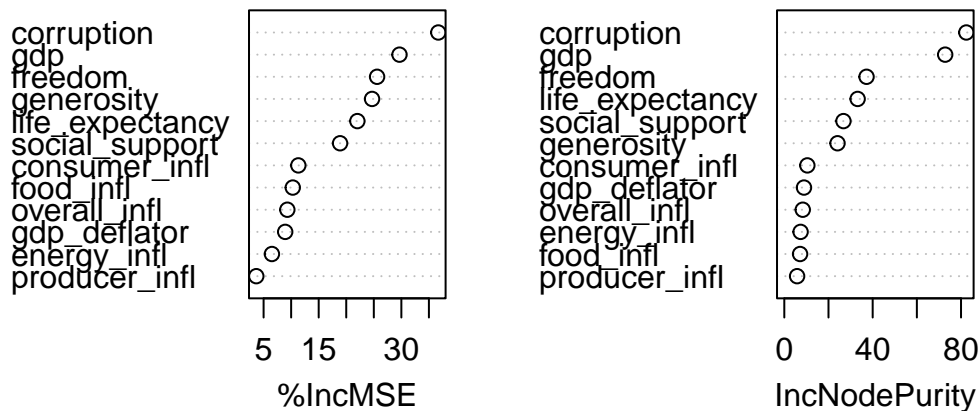
rf_results <- lm_train %>% predict(WHI_test)
data.frame(R2 = R2(rf_preds, WHI_test$happy_score),
           RMSE = RMSE(rf_preds, WHI_test$happy_score),
           MAE = MAE(rf_preds, WHI_test$happy_score))

```

	R2	RMSE	MAE
1	0.8116599	0.4536838	0.3538655

```
varImpPlot(rf1)
```

rf1



The Random Forest model outperformed the linear regression model, achieving a test  $R^2$  of 86.2% and an RMSE of 0.400. The most important predictors identified by the Random Forest were corruption, GDP, freedom, and social support. These findings reinforce the critical role that both economic and social factors play in determining national happiness.