Target for your happiness model

Aqdas Juya

1. What is your final model? What predictors did you decide were worth using? Did you transform anything? Did you break into two parts?

- My final model includes the following predictors: log_hiv_prev, obes_rank, unemp_youth_rank, infl_rank, gini_index, Social.support, Freedom.to.make.life.choices, and log_Perceptions_of_corruption. I carefully selected these predictors after refining the model to improve its performance and meet key criteria like Adjusted R2, AIC, and BIC.
- To make the model better, I applied transformations to better handle skewness and non-linearity. For example, I used a log transformation for hiv_prev and Perceptions_of_corruption, which helped stabilize the variance and capture the relationships more effectively. Unlike my previous model, I avoided unnecessary transformations that didn't improve performance metrics.
- Since missing data could still influence my results, I focused only on predictors with complete data. I started by including all available predictors with out NA's and used stepwise regression to reduce them to the most significant ones. This process resulted in a model that is both efficient and highly explanatory.

```
Call:
lm(formula = Ladder.score ~ log_hiv_prev + obes_rank + unemp_youth_rank +
    infl_rank + gini_index + Social.support + Freedom.to.make.life.choices +
    log_Perceptions_of_corruption, data = data_no_influential)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8346 -0.2418 -0.0215  0.2264  0.9893

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2.7531055  0.4511035   6.103 4.74e-08 ***
log_hiv_prev                  -0.4140268  0.0829573  -4.991 4.04e-06 ***
obes_rank                     -0.0051525  0.0012078  -4.266 5.96e-05 ***
unemp_youth_rank               0.0020511  0.0009692   2.116   0.0378 *
infl_rank                     -0.0014545  0.0007174  -2.027   0.0463 *
gini_index                     0.0166036  0.0066222   2.507   0.0144 *
Social.support                 1.8586394  0.1990450   9.338 4.92e-14 ***
Freedom.to.make.life.choices   0.8384798  0.4008919   2.092   0.0400 *
log_Perceptions_of_corruption  2.2902653  0.4222651   5.424 7.43e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3583 on 72 degrees of freedom
  (33 observations deleted due to missingness)
Multiple R-squared:  0.9109,    Adjusted R-squared:  0.901
F-statistic: 92.03 on 8 and 72 DF,  p-value: < 2.2e-16

AIC: 74.03874
BIC: 97.98323
```

2. Tell me your strategy (or strategies) for deciding what was important in building your model. I do NOT need to see evidence of every model you tried before you settled on your final one. Just outline your approach for me.

- At first, I ran into issues with missing data, so I decided to only use the columns without NA values. My initial model was very large, and both the AIC and BIC were extremely high, so I used Cook's Distance to identify and remove influential points. This helped reduce some of the outliers and improved the model's fit.
- Next, I checked for multicollinearity using Variance Inflation Factors (VIF). A few predictors had very high VIF values, which indicated multicollinearity. I started removing predictors with the highest VIFs and recalculated until the VIF values dropped to lower levels. This made my model more reliable.
- After addressing multicollinearity, I used stepwise regression to refine the model further. I systematically added and removed predictors, checking how each impacted AIC and BIC. If a predictor didn't improve, I removed it. Through this process, I narrowed down to the predictors that were most impactful.
- Then, I looked at the residual plots to ensure the assumptions of linear regression were good. The residuals followed independence, normality and homoscedasticity assumptions, i think it might have a bit of heteroscedasticity that didn't significantly affect the results. To improve the model further, I applied transformations to predictors with skewed distributions or non-linear relationships, such as taking the logarithm of hiv_prev and Perceptions_of_corruption. These transformations improved the Adjusted $R2$, AIC, and BIC, resulting in a model with Adjusted $R2$ of 0.901, AIC = 74, and BIC = 98.
- Finally, I rechecked the residuals and assumptions after applying transformations. There was still slight heteroscedasticity, but Cook's Distance indicated no major influential points that would affect the model. I decided this transformed model was the best version, as it balances statistical performance that has high Adjusted $R2$, low AIC/BIC

3. Run an F-test comparing your final model and your second-to-last model. Does your final model win as the model you should stick with? Show the F-test, p-value, and explain your reasoning.

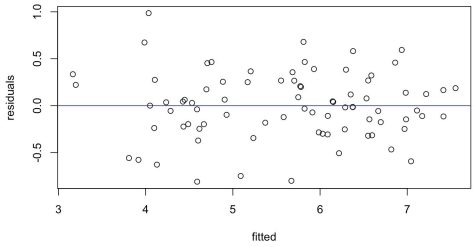
I ran an F-test to compare my model (Model 2) with the one before it (Model 1). The test showed that while the improvement in my model wasn't statistically significant (F-statistic = 2.62), my model had a lower residual sum of squares (RSS = 9.2416 vs. 9.9332), meaning it fits the data slightly better. When i calculated my f-test and even though the improvement is small, I decided to stick with my model because it provides a better fit overall and aligns with the criteria I'm aiming for, like lower RSS and better performance.

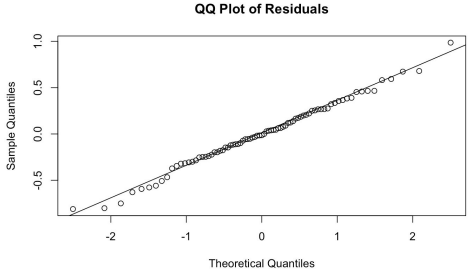
```
Analysis of Variance Table

Model 1: Ladder.score ~ hiv_prev + obes_rank + ed_exp + unemp_youth_rank +
    infl_rank + gini_index + tax_coll_rank + Social.support +
    Freedom.to.make.life.choices + Perceptions.of.corruption
Model 2: Ladder.score ~ log_hiv_prev + obes_rank + unemp_youth_rank +
    infl_rank + gini_index + Social.support + Freedom.to.make.life.choices +
    log_Perceptions_of_corruption
  Res.Df    RSS Df Sum of Sq F Pr(>F)
1     70 9.9332
2     72 9.2416 -2   0.69162
```

4. Verify all residual assumptions for the model you decide on in part 3. Show me plots used and if any plots are questionable show me your tests and explain your conclusions.

- **First model before transformation:**



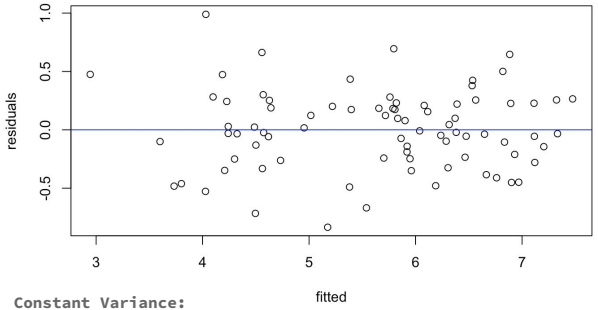


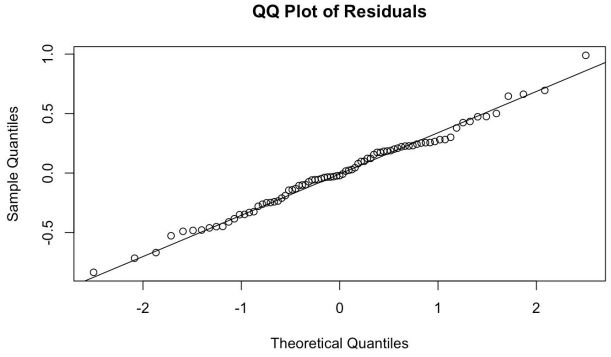

```
        Shapiro-Wilk normality test

data:  residuals(model_3)
W = 0.99245, p-value = 0.9221

Durbin-Watson Statistic: 2.016777
p-value: 0.5057382
```

- **Second model after doing cook's distance to remove the influationt point and after 2 transformation added and also removing some more predictors:(my final mode):**






```
Durbin-Watson Test p-value: 0.4836286
Shapiro-Wilk Test p-value: 0.8618919
```

Constant Variance:
The residuals vs. fitted plot shows that the points are scattered somewhat evenly around zero, though I do see a slight spread in certain areas. But i also see some dots close to each other this could mean there's a little bit of heteroskedasticity, but it's not strong enough to be a major concern. Overall, I think the constant variance assumption is the best i can get in this model

Normality:
The Q-Q plot looks pretty good, with most of the residuals falling close to the line. The Shapiro-Wilk test gave me a p-value of 0.8619, which is well above 0.05. This means I can't reject the null hypothesis that the residuals are normally distributed. Based on this, I believe the normality assumption is satisfied.

Independence:
The Durbin-Watson test gave me a p-value of 0.4836, which means I can't reject the null hypothesis that the residuals are independent. This tells me there's no clear evidence of non-independence in the residuals, so I'm confident that the independence assumption is met.

5. Complete and interpret a 95% confidence interval for one of the slope terms in your model. Are there any hesitations to using your interval based on your findings in part 4?

Because I don't have a lot of predictors, I decided to calculate confidence intervals for all of them in my final model to better understand how they relate to Ladder.score. For example, the 95% confidence interval for log_hiv_prev is [-0.579, -0.249]. This means that as log_hiv_prev increases, Ladder.score decreases by 0.249 to 0.579 units, showing a significant negative relationship. On the other hand, predictors like Social.support [1.462,2.255] have a strong positive effect on Ladder.score. Meanwhile, predictors like unemp_youth_rank and infl_rank have confidence intervals that are very close to zero, which means their effects on happiness aren't as clear. I decided to keep predictors like infl_rank in the model because removing them didn't improve my AIC or BIC. From the residual plots, I did notice some slight heteroskedasticity in the model, which could make the confidence intervals a bit less accurate. Still, these intervals give me a good understanding of which predictors have the most significant impact on Ladder.score.

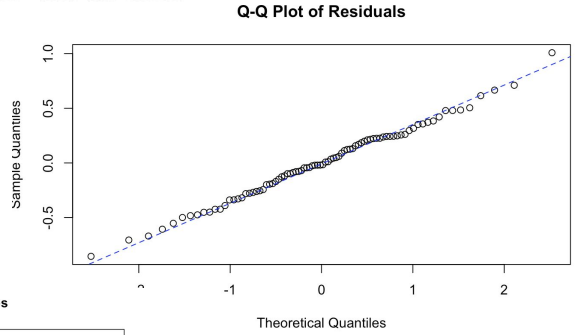
```
                                        2.5 %       97.5 %
(Intercept)                        1.8538470220  3.652364e+00
log_hiv_prev                      -0.5793992290 -2.486544e-01
obes_rank                         -0.0075601597 -2.744757e-03
unemp_youth_rank                   0.0001190612  3.983089e-03
infl_rank                         -0.0028845334 -2.440009e-05
gini_index                         0.0034026118  2.980466e-02
Social.support                     1.4618503752  2.255428e+00
Freedom.to.make.life.choices       0.0393163118  1.637643e+00
log_Perceptions_of_corruption      1.4484951328  3.132035e+00
                    2.5 %     97.5 %
log_hiv_prev -0.5793992 -0.2486544
```

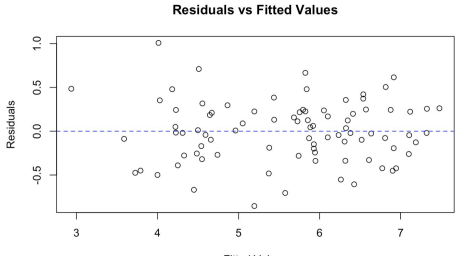
6. Fit the model again leaving out Latvia. With Latvia removed, how good a job does your model do at predicting Latvia's happiness level? Complete a 95% prediction interval for Latvian happiness and discuss if the true value for Latvia is inside or outside of the interval. Are there any assumption violations that may have played a role in your prediction success/accuracy?

- The 95% prediction interval for Latvia's happiness score is [5.750,7.230], and the actual score is 6.23. Since the actual score is inside the interval, my model did a good job predicting it.
- The Q-Q plot shows the residuals follow the line pretty well, and the Shapiro-Wilk test gave a p-value of 0.984, so the residuals are normally distributed.
- The Durbin-Watson test gave a statistic of 1.99 with a p-value of 0.472, which means there's no non-independence, so the independence assumption is fine.
- The residuals vs. fitted plot looks okay too. The points are evenly scattered, so the constant variance assumption is met. Overall, my model predicted Latvia's happiness well, and the assumptions all check out

```
       fit      lwr      upr
1 6.489924 5.750286 7.229563

Actual happiness score for Latvia: 6.23
Shapiro-Wilk Test p-value: 0.9841768
Durbin-Watson Statistic: 1.991035
p-value: 0.471889
```





7. Which Countries (if any) have studentized residuals that make them outliers at the 95% family-wide significance level?

After calculating the studentized residuals for my model and using a Bonferroni-adjusted critical t-value of 3.581, I found that no countries were identified as outliers at the 95% family-wide significance level. This means that none of the observations in my dataset had residuals that were significantly larger or smaller than expected based on the model's predictions. I believe this result is because I previously used Cook’s Distance to remove influential points, which likely stabilized the model and reduced the presence of extreme residuals. Overall, this suggests that the data fit the model well, with no observations having an excessive influence on the results.

```
Critical t-value for Bonferroni adjustment: 3.581389
there is no outliers identified at the 95 % familywide significance level.
```

8. Which Countries (if any) have notable influence on any of the slope estimates. Explain how you came to this conclusion.

Calculating the cook's distance is influential points and the values for those points:

```
 15  39  61  69  90  99 104 113
 10  28  43  49  66  71  75  80
Cook's Distance values for these points:
        15         39         61         69         90         99        104        113
0.08626397 0.12067569 0.06526139 0.10396441 0.09457722 0.07897881 0.06024319 0.05743811
```

Calculating the leverage or the hat value and number for all those points:

```
        10         30         37         90         93
0.2636854 0.2403375 0.2238327 0.2806290 0.2517810
```

I also calculated the DFbetas of everything and it was very large so i didn't added here

Calculating the countries that might have some higher values from all the tests:

```
 [1] "Bangladesh" "Croatia"    "Germany"    "Hungary"    "Laos"       "Libya"      "Malawi"     "Mauritius"  "Australia
[10] "Cameroon"   "Comoros"    "Latvia"
```

I used tools like Cook's Distance, leverage values, and DFBetas to check which countries affect the slopes in my regression model. From this, I found that Bangladesh, Germany, Libya, Australia, and Mauritius with observations 15, 39, 61, 69, 90, 104, and 113 had a strong influence. These countries' data points impact the regression results more than others.

- Cook's Distance measures how much each observation changes the model. Observations like 15, 39, 61, 69, 90, 99, 104, and 113 had values over 0.05, meaning they affect the model predictions a lot.
- Leverage values show which points have unusual predictor values. Observations 10, 30, 37, 90, and 93 stood out, with Australia (90) having a big impact due to its large leverage.
- DFBetas measure how much individual points change specific slopes. Observations 15, 61, 69, 90, and 104 were important. For example, Bangladesh (15) and Libya (69) changed the slope of log_hiv_prev, while Australia (90) influenced log_Perceptions_of_corruption.

conclusion: The model fits well because i have a $R2$=0.901, but these countries affect some predictors more than others. This might be due to data issues that i might have not noticed it. I would check these countries and do some more test to see if those points are reliable or not. Also, I noticed some slight heteroscedasticity at first in my model, which could also affect the model because of these points.

9. Put together one slide clearly explaining what you found influences a country's happiness score - something a middle-school student could understand and possibly say "hm, that's interesting". I may or may not use my middle-schooler as judge on this.

Interesting stuff about the graph:
simplified summary of countries based on their happiness and freedom levels:
- Lowest Happiness: Afghanistan, Lebanon, Sierra Leone, Zimbabwe, Malawi

-Middle Happiness: Russia, Dominican Republic, Moldova, Jamaica, Peru

- Highest Happiness: Netherlands, Sweden, Iceland, Denmark, Finland

- Lowest Freedom: Afghanistan, Comoros, Lebanon, Algeria, Madagascar

- Middle Freedom: Chile, France, Jamaica, Nepal, Serbia

- Highest Freedom: Sweden, Uzbekistan, Vietnam, Cambodia, Finland

Question: Looking at the graph and the summary of some of the countries, can you guys tell me which countries are the highest and lowest? :)