**What Truly Makes us Happy: An Investigation of the World Happiness Report**

1. <https://www.jstor.org/stable/2093503?seq=1>

2. <https://s3.amazonaws.com/happiness-report/2019/WHR19.pdf>

3. <https://journals.sagepub.com/doi/pdf/10.1177/0146167211407075>

4. <https://www.tandfonline.com/doi/abs/10.2747/0272-3638.32.6.871>

5. <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1559-1816.2010.00671.x?casa_token=3GKT7Unvgn0AAAAA%3ATJR3K4u37p7ZfdzsGJLeveS7OOGivrf-dx0CDZAFNUIJR3fUNBIEYz4KDEOz8JjcrMijWQ3VD6m3mdo>

1. This is what the report declared what mattered: <https://www.kaggle.com/unsdsn/world-happiness>
2. The world happiness report is a survey, every 2 years, of people in various countries to see their self-reported levels of happiness.
3. However, because of the following literature, we wanted to add different variables to see what’s a better predictor in predicting happiness, what other metrics can also be used to predict happiness, and finally, to see if there’s a way to improve the way we understand what makes us happy people.
   1. Population density literature // Behavioral sink theory
   2. Life expectancy literature
4. To ensure the quality of our data and our model, we ran diagnostics
   1. Multicollinearity
      1. Ran into it with quality of life variables: Life expectancy, fertility rate, percent electricity access - we dropped fertility rate to preserve r^2 but reduce the VIF/Correlation matrix
   2. Autocorrelation
      1. No temporal autocorrelation
      2. But we may have spatial autocorrelation
         1. Use leaflet to plot happiness ranking to show the spatial
   3. Heteroskedasticity
   4. Omitted Variable Bias
5. We found that --
   1. Life expectancy is a stronger predictor than fertility rate when checking for quality of life metrics; Electricity access, a measure of infrastructure and also arguably quality of life, also ended up being a strong predictor
   2. Population density variables, while not the strongest predictors, became significant ones which help contribute to the strong predictive ability of our model
   3. Strength of legal rights (a variable from the World Bank Bank Open Dataset) is actually not a significant variable, even though the happiness report said that it should be, as it falls under the 4th most powerful category of variables, freedom and legal rights.

**Project Notes 10/19/19**

***What we have so far:*** a linear regression model for happiness ranking based on: gdp, life expectancy, unemployment, fertility, pop density, urban population, female labor, legal rights, electricity access, and air pollution

***What we’re trying to do:*** Figure out if the significance of the variables are real or if they’re skewed by multicollinearity by looking at the data with and without certain variables. If so: which variables should we keep?

***Notes as we go:***

Ran proc corr with data with all variables

* Results --

[image of the corr table]

* + GDP has strong correlation with life expectancy (.655), urban population (.61625)
  + Life expectancy has strong correlation with gdp, Fertility rate (.85419), urban population (.70454), electricity access (.83781)
  + Fertility rate with life expectancy, percent electricity access (.86296)
  + Urban pop has strong correlation with GDP, life expectancy, Fertility rate, and percent electricity access (.64142)
  + Electricity access has strong correlation with life expectancy, fertility, urban pop
* Issues --
  + We have a group of correlated variables (seemingly): fertility, life expectancy, urban population, and electricity access
    - *Potential solution*, see what the data looks like without each
* Suggestions Decided
  + Talk to kevin about the strong correlations // the stuff on alternative metrics for multicollinearity → figure out what to do about the variables
  + Drop broadband; drop legal rights

**UMich Multicollinearity Notes 10/20/19** -- file:///C:/Users/saipr/Desktop/day3\_finan\_collin.pdf

* Multicollinearity -- Multiple predictors in a multiple regression model that are highly correlated with each other
  + Have adverse effects:
    - Their significance tests (SEs tend to be inflated)
    - They may seem like they’re significant predictors, but they also seem to fail the significance test; *similar to life expectancy/fertility in the model*
  + Can be tested on using additional options to proc reg
    - Vif tol collin
      * The tolerance option is the proportion of variance that’s not explained by the other predictors
      * VIF (Variance Inflation Factor) if a factor by which the variance of the estimated coefficient is multiplied by b/c the multicollinearity
  + A good check for multicollinearity problem:
    - Correlation matrix thresholds is above .65
    - Condition Index is greater than 30
    - Eigenvalue: decimal has more than 3 decimal points
    - If VIF exceeds 5 or 10, implies associated regression coefficients are poorly estimated because of multicollinearity

Code:

/\* refit model, and request collinearity diagnostics \*/

proc reg data = cars2;

model mpg = weight year engine horse accel cylinder / vif tol collin;

run;

Quit;

**Project Noes 11/03/19**

***What we have so far:*** a linear regression model with a very clean dataset and a solid enough set of variables (though they seem to be susceptible to multicollinearity)

* Note — is the world happiness report flawed because, as we’ve seen, some of their variables have multicollinearity? What does this mean for the significance of their report?

***What we’re trying to do:*** Figure out the best pathway to correct to multicollinearity; figure out if we suffer from omitted variable bias, heteroskedasticity, or autocorrelation.

***Notes as we go:***

* *With ALL variables* — R^2 is .7600
  + Life expectancy has a P value of .2976 / VIF of 7.454
  + Fertility rate has a P value of .9401 / VIF of 5.981
  + Electricity Access has P value of .0338 / VIF of 6.595
* *With Life Expectancy Omitted* — R^2 is .7602
  + .
  + Fertility rate has a P value of .6565 / VIF of 5.084
  + Electricity access has a P value of .1805 / VIF of 5.541
* *With Fertility Rate Omitted* — R^2 is .7619
  + Life Expectancy has a P value of .3012 / VIF of 6.2122
  + .
  + Electricity access has a P value of .23306 / VIF of 4.291
* *With Electricity Access Omitted* — R^2 is .7506
  + Life Expectancy has a P value of .0682 / VIF of 6.44500
  + Fertility Rate has a P value of .0985 / VIF of 4.08482
  + .
* *With JUST life expectancy* — R^2 is .7469
  + Life Expectancy has a P value of .0001 / VIF of 2.92576
  + .
  + .
* *With JUST Fertility Rate* — R^2 is .7459
  + .
  + Fertility Rate has a P value of .0001 / VIF of 1.26511
  + .
* *With JUST Electricity Access* — R^2 is .7617
  + .
  + .
  + Electricity Access has a P value of .001 / VIF of 1.36635

***Decision:*** As the highest r^2 values exist in case 3 and case 7 (no fertility, just electricity), we examine VIFs and P values; as we see in the individual breakdowns, life expectancy is clearly significant and thus omitting it loses a quality statistic on for consideration; as Fertility rate is inherently included in Life Expectancy —> we choose CASE 3 — FERTILITY RATE OMITTED

**Project Notes 11/04/19**

***What we’ve done so far***: We tested and found multicollinearity in this model, and ultimately decided that the model with Fertility Rate omitted is the best balance of reducing multicollinearity, increasing P value, and increasing the R^2 value. This is now the official model.

***What we’re doing now***: We want to check for autocorrelation by doing the DW test

***Notes as we go***:

* So we typed in /dw
  + Durbin Watson D = 1.377
  + Rho = .310
  + Table Statistic = 1.35 for DF of 125 (134 - 9) and K of 9; upper bound of 1.76
* As our test statistic is only marginally higher than the lower bound, we have inconclusive data. Since time is not an issue here, it dosen’t matter.
  + We may suffer from spatial correlation since we have geographic data -- <https://scholar.google.com/scholar?q=autocorrelation+in+geographic+data&hl=en&as_sdt=0&as_vis=1&oi=scholart>

**Project Notes 11/17/19**

***What We’ve Done So Far:*** Corrected for multicolinearity, found a bit of a quirk in the autocorrelation part of the dataset

***What to do now***: OVB and Heteroskedasticity

***Project Notes***

* OVB
  + SSEr = 54594
  + SSEu = 54116
  + J = 2
  + N = 129
  + K = 7
  + Test statistic = .539
  + Table statistic = 1.43
  + As our test statistic is lower than our table statistic we DO NOT suffer from OVB
* Heteroskedasiticty
  + Based on the residuals, we have to check heteroskedasticity for these variables (funnel like residual plots)
    - GDP per capita
    - Life expectancy
    - Unemployment rate
    - Air pollution
    - People per square meter
  + We use the /spec to check fo heteroskedasticity
    - As our pr > chisquared is greater than .01, we HAVE heteroskedasticity
  + We corrected it with /white