Final Exam

May 17, 2019

- My name is Matthew Rodgers.
- This is the final exam.

Importing

```
In [98]: import pandas as pd
         import numpy as np
         import statsmodels.api as sm
         import statsmodels.formula.api as smf
         import matplotlib.pyplot as plt
         import matplotlib
         import warnings
         warnings.filterwarnings('ignore')
         url = "https://raw.githubusercontent.com/ajr348/happiness/master/happiness.csv"
```

1 Cleaning & Organizing

```
In [104]: data = pd.read_csv(url)
          data.set_index("Year", inplace=True)
          year2018 = data.loc[2018]
          year2018["Percentiles"] = year2018["Log GDP per capita"].rank(ascending=False)
          year2018.set_index("Country name", inplace=True)
          print(year2018.loc["Argentina"])
```

```
Life Ladder                                    5.792797
Log GDP per capita                             9.809972
Social support                                 0.899912
Healthy life expectancy at birth              68.800003
Freedom to make life choices                   0.845895
Generosity                                    -0.206937
Perceptions of corruption                      0.855255
Positive affect                                0.820310
Negative affect                                0.320502
Confidence in national government              0.261352
Democratic Quality                                  NaN
Delivery Quality                                    NaN
Standard deviation of ladder by country-year   2.472559
```

```
Standard deviation/Mean of ladder by country-year            0.426833
GINI index (World Bank estimate)                                  NaN
GINI index (World Bank estimate), average 2000-16            0.460938
gini of household income reported in Gallup, by wp5-year     0.405356
Most people can be trusted, Gallup                                NaN
Most people can be trusted, WVS round 1981-1984              0.270073
Most people can be trusted, WVS round 1989-1993              0.223553
Most people can be trusted, WVS round 1994-1998              0.170844
Most people can be trusted, WVS round 1999-2004              0.150154
Most people can be trusted, WVS round 2005-2009              0.174058
Most people can be trusted, WVS round 2010-2014              0.193531
Percentiles                                                 45.000000
Name: Argentina, dtype: float64
```

- Argentina is in the 45th percentile meaning that Argentina 45% of countries in Log GDP per capita

2 Descriptive Statistics

```
In [105]: data["Log GDP per capita"].median()
```

```
Out[105]: 9.406206131
```

- The median Log GDP per capita is 9.406206131

```
In [106]: data["Log GDP per capita"].loc[2016].median()
```

```
Out[106]: 9.526675224
```

- The meadian Log GDP per capita in 2016 is 9.526675224

```
In [102]: data["Perceptions of corruption"].astype(float)
          from scipy.stats import iqr
          iqr(data["Perceptions of corruption"])
```

```
Out[102]: nan
```

```
In [103]: data["Freedom to make life choices"].loc["United Kingdom"].std()
```

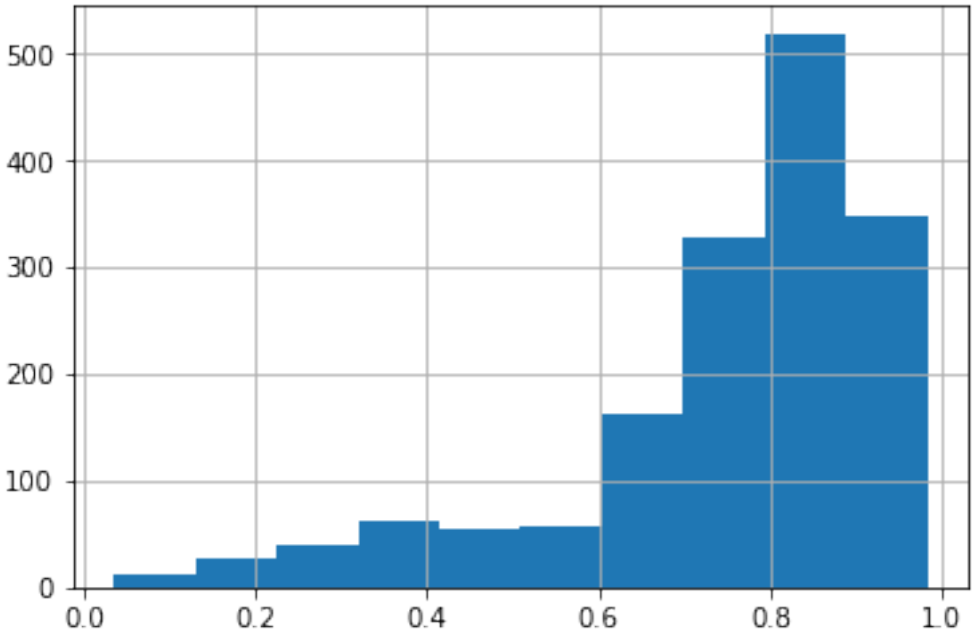
```
Out[103]: 0.04526501059465575
```

- The standard deviation of Freedom to make life choices in the UK is 0.04526501059465575

3 Graphing

```
In [82]: data["Perceptions of corruption"].hist()
```

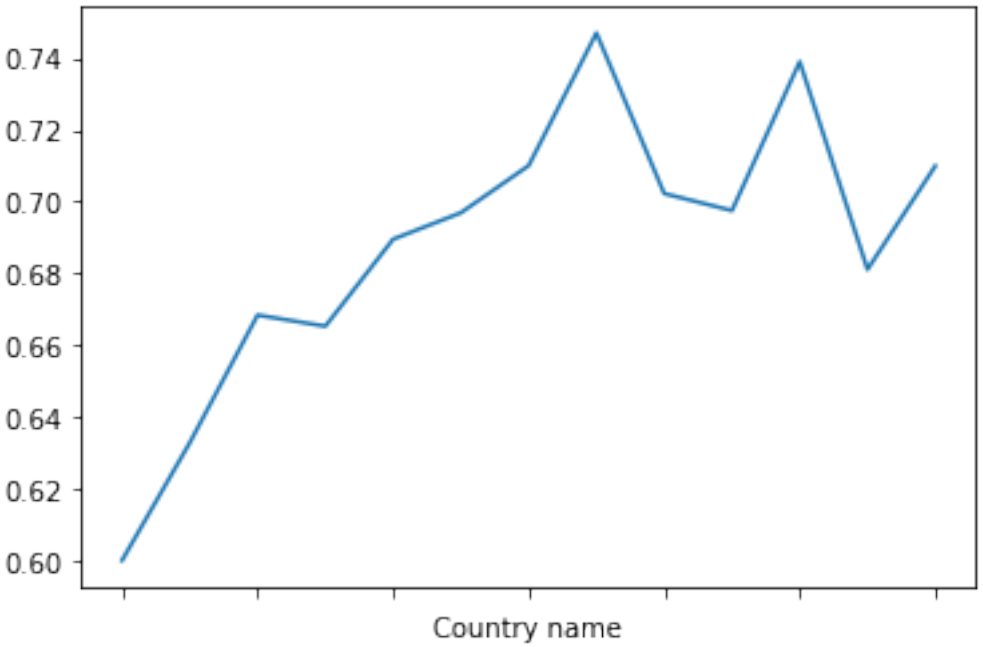
```
Out[82]: <matplotlib.axes._subplots.AxesSubplot at 0x2b8a153d7b00>
```



- The histogram of perceptions of corruption shows the data skewed left. This means that most of the values are in the upper range.

```
In [91]: data["Perceptions of corruption"].loc["United States"].plot()
```

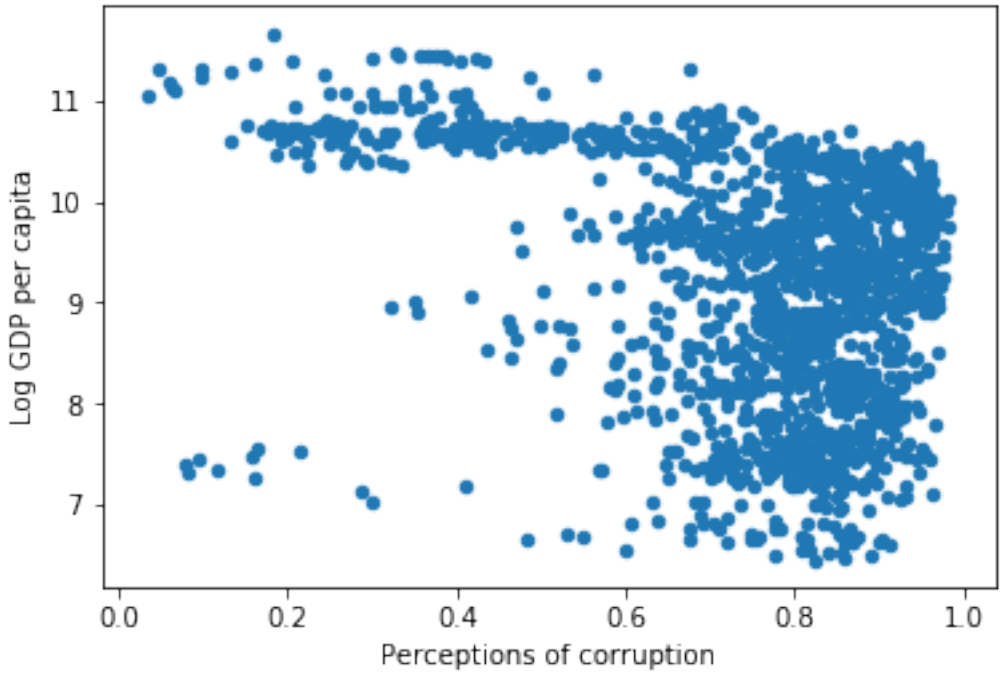
```
Out[91]: <matplotlib.axes._subplots.AxesSubplot at 0x2b8a16c45978>
```



- Over time perceptions of corruption has increased in the United States

```
In [92]: data.plot.scatter("Perceptions of corruption", "Log GDP per capita")
```

```
Out[92]: <matplotlib.axes._subplots.AxesSubplot at 0x2b8a16c40f98>
```



- The scatter plot shows a negative relationship

4 Hypothesis formation

- Log GDP Per Capita = a + b*Perception of Corruption
- H0: There is no relationship between Log GDP Per Capita and Perceptions of corruption.
- HA: There is a negative association between Log GDP Per Capita and Perceptions of corruption.

5 Regression

```
In [107]: data.rename(columns={"Perceptions of corruption": "corruption"}, inplace = True)
          data.rename(columns={"Log GDP per capita": "gdp"}, inplace = True)
          results = smf.ols("gdp ~ corruption", data = data).fit()
          results.summary()
```

```
Out[107]: <class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                    gdp   R-squared:                       0.116
Model:                            OLS   Adj. R-squared:                  0.116
Method:                 Least Squares   F-statistic:                     208.0
Date:                Fri, 17 May 2019   Prob (F-statistic):           2.15e-44
Time:                        13:33:57   Log-Likelihood:                -2410.0
No. Observations:                1581   AIC:                             4824.
Df Residuals:                    1579   BIC:                             4835.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     10.8245      0.117     92.475      0.000      10.595      11.054
corruption    -2.1766      0.151    -14.423      0.000      -2.473      -1.881
==============================================================================
Omnibus:                      121.316   Durbin-Watson:                   0.224
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              115.624
Skew:                          -0.603   Prob(JB):                     7.81e-26
Kurtosis:                       2.451   Cond. No.                         8.53
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
"""
```

6 Interpretation & diagnostics

- The Coefficient is -2.1766. This tells us that with a one unit change in Perceptions of corruption the gdp changes by -2.1766.
- The P value of .000 means that there is roughly a 0% probability of seeing this coefficient if the null hypothesis of no relationship is actually the case.
- The confidence interval is 95% ($p < .05$). The confidence interval means that if we conducted the same experiment many times the percentage of confidence intervals that contained the true population mean would be 95%.
- The R2 and Adj. R2 shows how well the regression line fits the data. The R^2 here shows that approximately .116 of the observed variation can be explained by the model's inputs.
- The Prob(F-Statistic) tells us that there is a 2.15e-44 probability that the null hypothesis in the regression model cannot be rejected
- We reject the null hypothesis because the p-value in this model is < .05.
- The model satisfies the major assumptions of OLS regression. There is a reasonable sample size, a linear relationship, low multicollinearity, minimal outliers, and homoscedasticity.
- A bias present that concerns me is response bias because there is no objective way to rate your perceptions of corruption

7 Conclusion

- I am not very confident that I understand the relationship between GDP and corruption because of the low r^2 value given by the regression.