STAT 632, HW 5
Due: Tuesday, March 24

Reading: Chapter 6, pp. 151–162, from *A Modern Approach to Regression*.

The Human Development Index (HDI) is an index developed by the United Nations to assess the development of a country. The HDI is calculated by combining three indicators: life expectancy at birth, average education, and gross national income per capita.[1] The following exercises will investigate the relationship between HDI and other variables on national demographics, connectivity (internet usage), and employment.[2]

To load the data into R run the following command:

```
hdi = read.csv("https://ericwfox.github.io/data/hdi2018.csv")
```

Variable descriptions:

- `hdi_2018`: HDI for the year 2018
- `median_age`: Median age (years) in 2015
- `pctpop65`: Percent of population 65 and older in 2018
- `pct_internet`: Percent of population that uses the internet in 2017-2018
- `pct_labour`: Percent of country's working-age population that engages actively in the labour market, either by working or looking for work in 2018

Exercise 1

(a) Fit a multiple linear regression model with `hdi_2018` as the response, and the other four variables as predictors.

(b) Using the model fit in (a), is there evidence of a relationship between `hdi_2018` and at least one of the predictor variables? Write the null and alternative hypotheses, report the F-test statistic and p-value, and state your conclusion.

(c) Using the model fit in (a), which predictor variables are statistically significant according to the individual t-tests?

[1] `http://hdr.undp.org/en/content/human-development-index-hdi`
[2] Date obtained from `http://hdr.undp.org/en/data`

(d) Fit a reduced model with `median_age` and `pct_internet` as predictors. Use the `anova()` function to conduct a partial F-test that compares this reduced model with the full model specified in (a). Make sure to write the null and alternative hypotheses, report the p-value, and state your conclusion.

(e) According to the adjusted-R^2, how does the full model in (a) compare with the reduced model in (d)? Is this consistent with your conclusion for the partial F-test?

Exercise 2. For this exercise, consider the regression model with `hdi_2018` as the response, and `median_age` and `pct_internet` as predictors.

(a) Make a scatterplot matrix for the three variables. Describe the associations between the variables in the scatterplot matrix.

(b) Make a plot of the residuals versus fitted values, and a QQ plot of the standardized residuals.

(c) Make a plot with the leverage values (h_i) on the x-axis, and standardized residuals (r_i) on the y-axis. Identify any points (countries) that have high standardized residuals or leverage.

(d) Based on the scatterplot matrix and model diagnostics, do the assumptions for MLR appear adequately satisfied? Can you think of any ways in which the model might be improved to better fit the data?