PM592: Regression Analysis for Data Science

Name:
Flemming
Wu

HW4

Multiple Linear Regression, Regression Diagnostics

Instructions

- Answer questions directly within this document.
- Upload to Blackboard by the due date & time.
- Clearly indicate your answers to all questions.
- If a question requires analysis, attach all relevant output to this document in the appropriate area. Do not attach superfluous output.
- There are 3 questions and 30 points possible.

Question 1 [7 points]

Read the article by Liang et al. (2020) on Blackboard.

1a. [1 point] Where did the authors obtain their data for this study? How many countries were included in the analysis? Were any countries excluded and, if so, why?

The authors obtained data for their study from a variety of sources: the Worldometer website, the Worldwide Governance Indicators, World Development Indicators, and Logistics Performance Indicators database. There were 169 countries included in the analysis, and ones without publicly-available data COVID-19 mortality rate were excluded. These countries were excluded because the outcome or dependent variable was COVID-19 mortality rate in this analysis, so without that data the researchers cannot perform a regression.

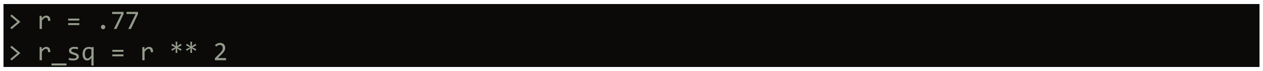
1b. [2 points] Suppose we wanted to lower Covid-19 mortality rate by providing countries with more test kits, but resources are limited. Based on Figure 1g-i, should these test kits be given to countries with high, moderate, or low percentage of aged persons? Why?

Based on Figure 1g-I, it looks like countries with low percentage of aged persons would stand to benefit the most from more test kits. The association is the strongest for test number per 100 people (log) vs. COVID-19 mortality rate percent (log) for countries with low percentage of aged persons, with the lowest p-value ($p < 0.001$) compared to moderate percentage of aged persons ($p=0.006$) and high percentage of aged persons ($p=0.335$). Additionally, the r value, which gives the correlation between test number per 100 people (log) and COVID-19 mortality rate percent (log), is the furthest from 0 of the three categories ($r=-0.67$) indicating the strongest correlation.

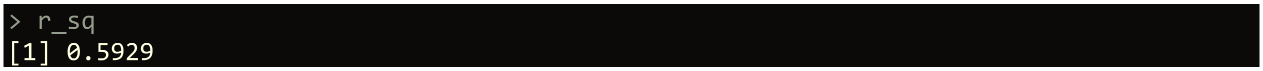
1c. [2 points] In which panel in Figure 1 do you believe influential observations may be affecting the regression line the most? Why?

I believe Figure 1d contains influential observations that affect the regression line the most out of the panel. There are three observations on the left side, which are further from the mean of all X values than the other observations and seem to cause the statistical software to flatten the slope of the regression line as it minimizes the sum of square residuals. Without those three points, I think the regression line in Figure 1d would have a steeper negative slope.

1d. [2 points] In Figure 2 the authors correlate the observed Covid-19 mortality rate with the model-predicted mortality rate. Based on the results in this figure, can you (roughly) re-create the model R-squared value they report in the caption of Table 2?



```
> r = .77
> r_sq = r ** 2
```



```
> r_sq
[1] 0.5929
```

The authors reported an R-squared value of 0.58, which roughly matches the R-squared value I calculated from their provided r value. The slight difference is likely due to rounding of the r value when reported on Figure 2.

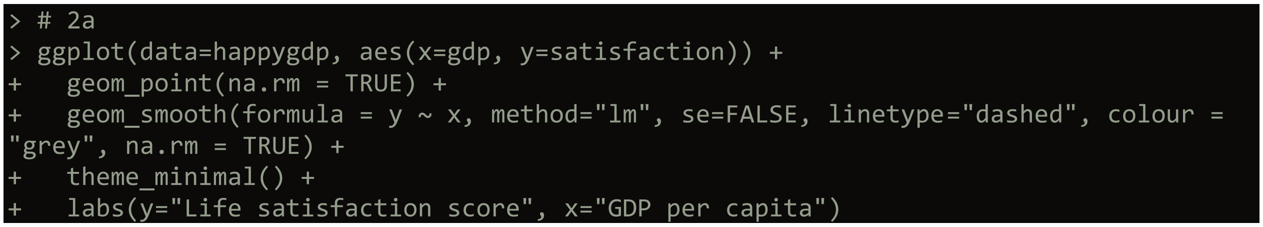
Question 2 [12 points]

Read the article located at: https://www.visualcapitalist.com/relationship-money-happiness/

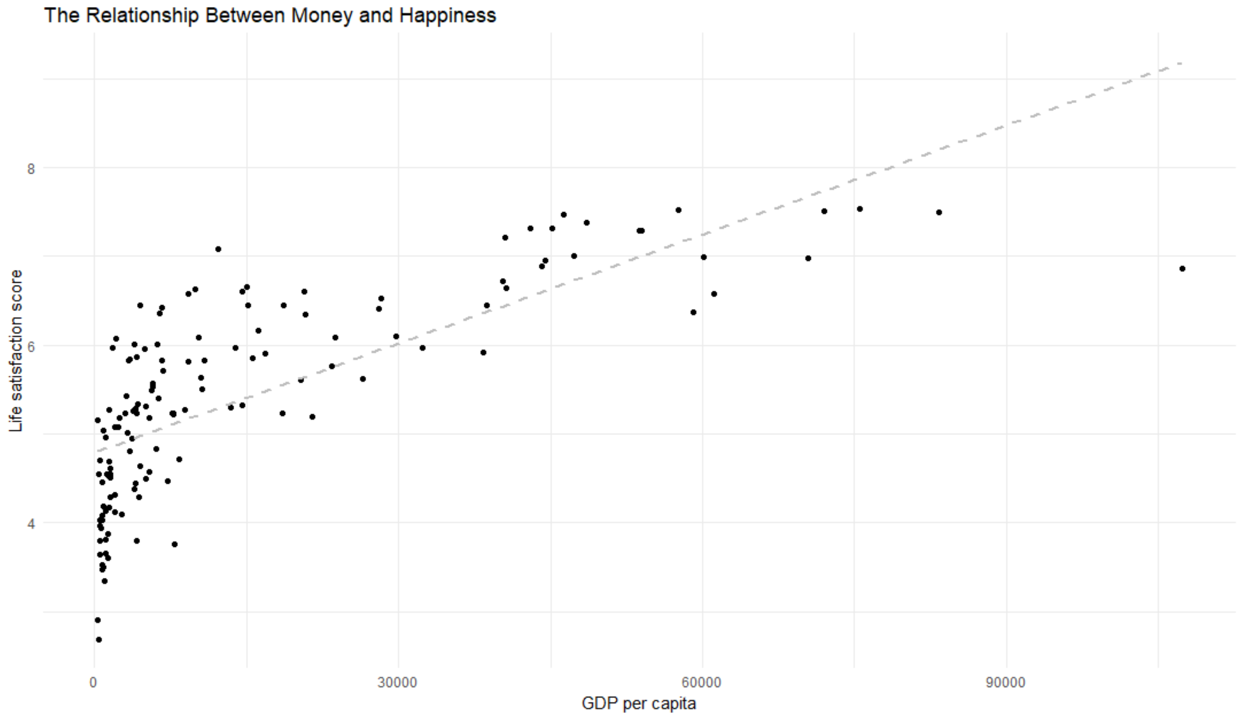
The happygdp.csv file on Blackboard contains data downloaded from the sources in the article – namely, GDP per capita (PPP; 2017 international dollars) and satisfaction index. Note that this downloaded data may not exactly match the data used in the figure.

Use this data set and remove any countries that are missing data.

2a. [1 point] Re-create the first figure in the article using ggplot. Make sure your figure uses a point and smooth geom.

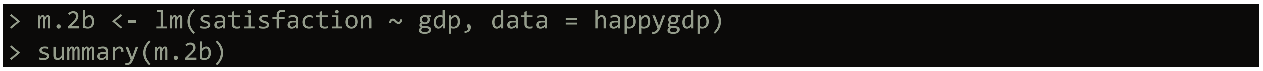


```
> # 2a
> ggplot(data=happygdp, aes(x=gdp, y=satisfaction)) +
+   geom_point(na.rm = TRUE) +
+   geom_smooth(formula = y ~ x, method="lm", se=FALSE, linetype="dashed", colour =
"grey", na.rm = TRUE) +
+   theme_minimal() +
+   labs(y="Life satisfaction score", x="GDP per capita")
```

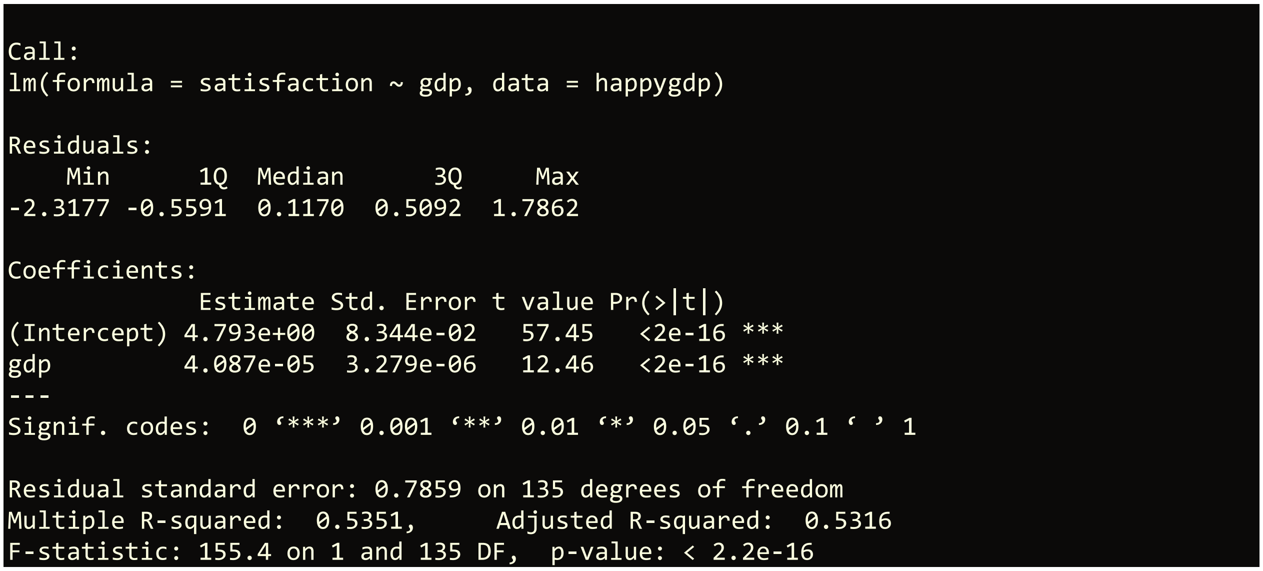


The Relationship Between Money and Happiness

2b. [2 points] Run a regression model of satisfaction on GDP. Provide the regression equation for the relationship between satisfaction and GDP.



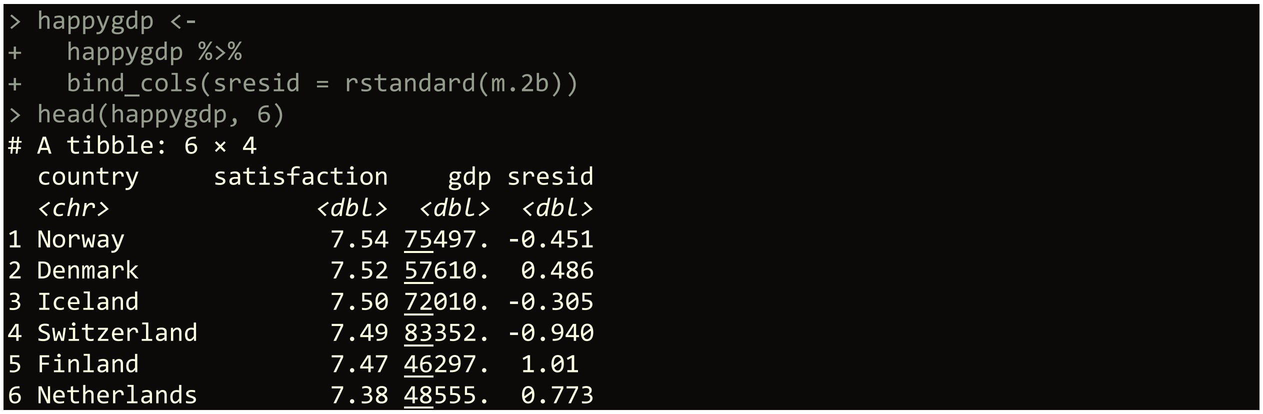
```
> m.2b <- lm(satisfaction ~ gdp, data = happygdp)
> summary(m.2b)
```



```
Call:
lm(formula = satisfaction ~ gdp, data = happygdp)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3177 -0.5591  0.1170  0.5092  1.7862 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.793e+00  8.344e-02   57.45   <2e-16 ***
gdp         4.087e-05  3.279e-06   12.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7859 on 135 degrees of freedom
Multiple R-squared:  0.5351,	Adjusted R-squared:  0.5316 
F-statistic: 155.4 on 1 and 135 DF,  p-value: < 2.2e-16
```

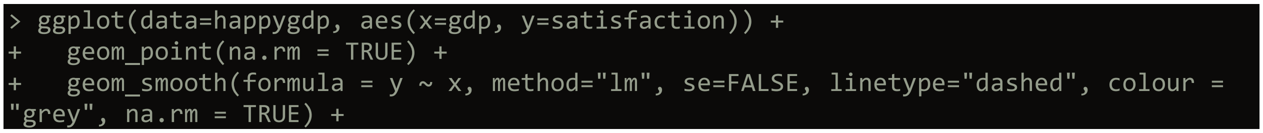
$$\hat{Y} = 4.79 + 0.000041X$$

2c. [2 points] Add the residuals from this model to your data set. Print out the first 6 observations to ensure the column has been added correctly.
Hint: data %>% bind_cols(sresid = rstandard(model))

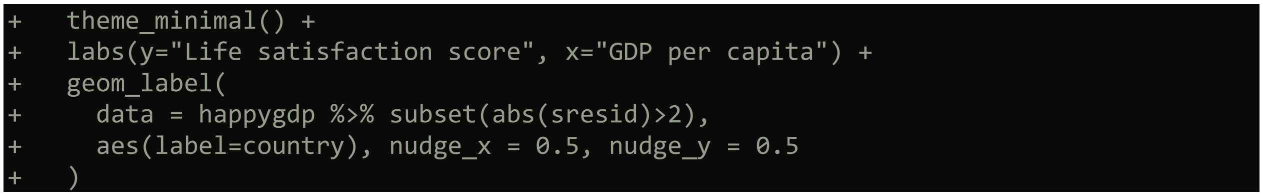


```
> happygdp <-
+   happygdp %>%
+   bind_cols(sresid = rstandard(m.2b))
> head(happygdp, 6)
# A tibble: 6 × 4
  country     satisfaction    gdp sresid
  <chr>              <dbl>  <dbl>  <dbl>
1 Norway              7.54 75497. -0.451
2 Denmark             7.52 57610.  0.486
3 Iceland             7.50 72010. -0.305
4 Switzerland         7.49 83352. -0.940
5 Finland             7.47 46297.  1.01 
6 Netherlands         7.38 48555.  0.773
```

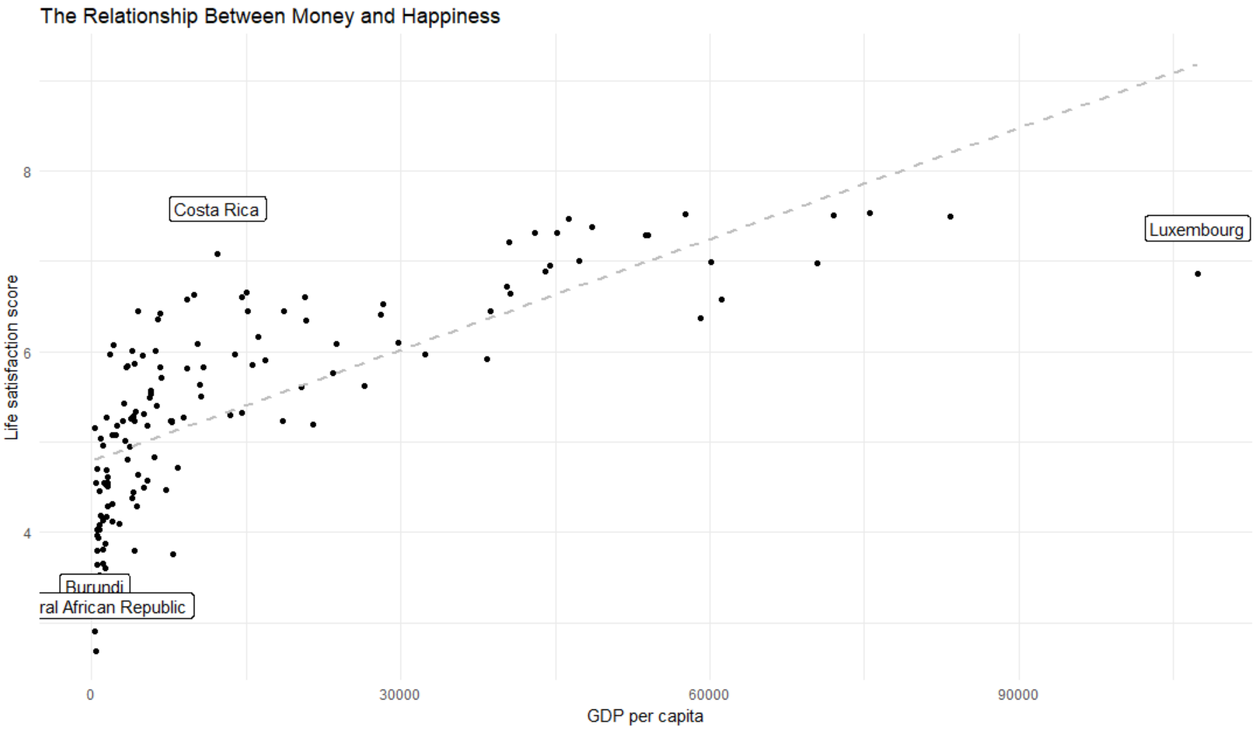
2d. [2 points] Update your figure from 2a to add labels for all countries that have a standardized residual that is greater than 2 in magnitude. Compare the outliers in your model/data to the outliers in the article.
Hint: geom_label(data=data %>% subset(abs(sresid)>2), aes(label=country))



```
> ggplot(data=happygdp, aes(x=gdp, y=satisfaction)) +
+   geom_point(na.rm = TRUE) +
+   geom_smooth(formula = y ~ x, method="lm", se=FALSE, linetype="dashed", colour = 
"grey", na.rm = TRUE) +
```



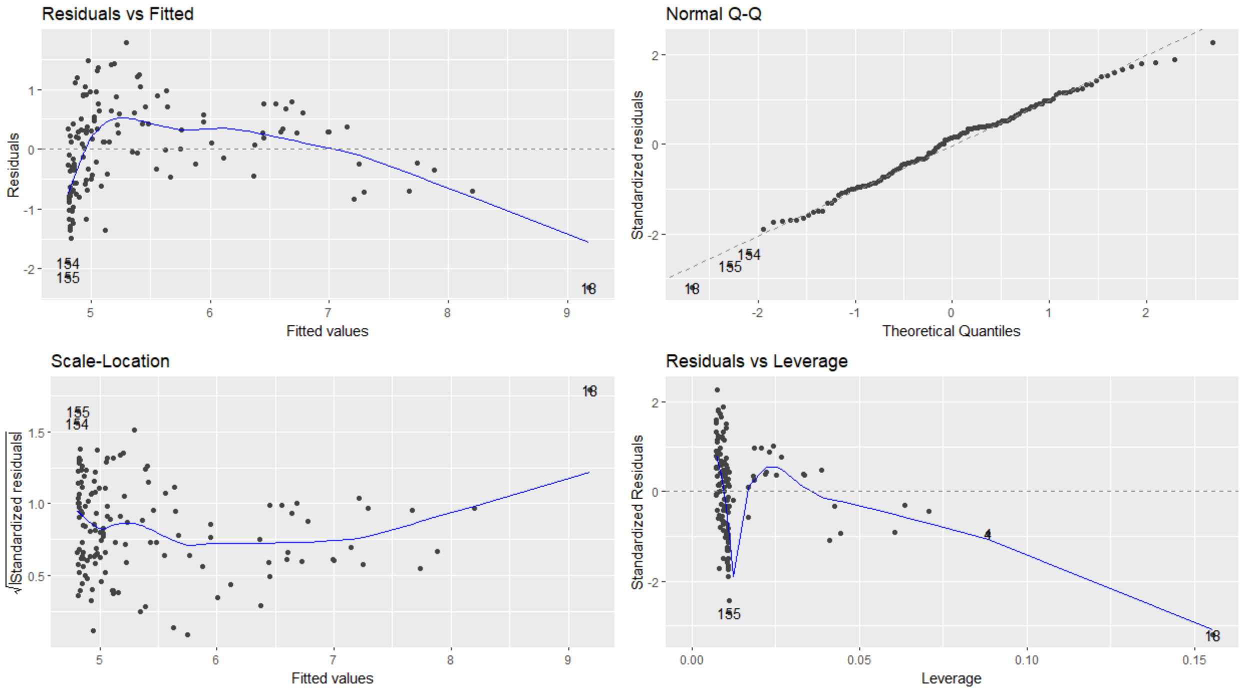
```
+   theme_minimal() +
+   labs(y="Life satisfaction score", x="GDP per capita") +
+   geom_label(
+     data = happygdp %>% subset(abs(sresid)>2),
+     aes(label=country), nudge_x = 0.5, nudge_y = 0.5
+   )
```



2e. [2 points] Did the authors check the assumptions of linear regression? Assess these assumptions for your model.

It was not stated whether the authors checked the assumptions of linear regression or not.

Assessing the assumptions on my own:



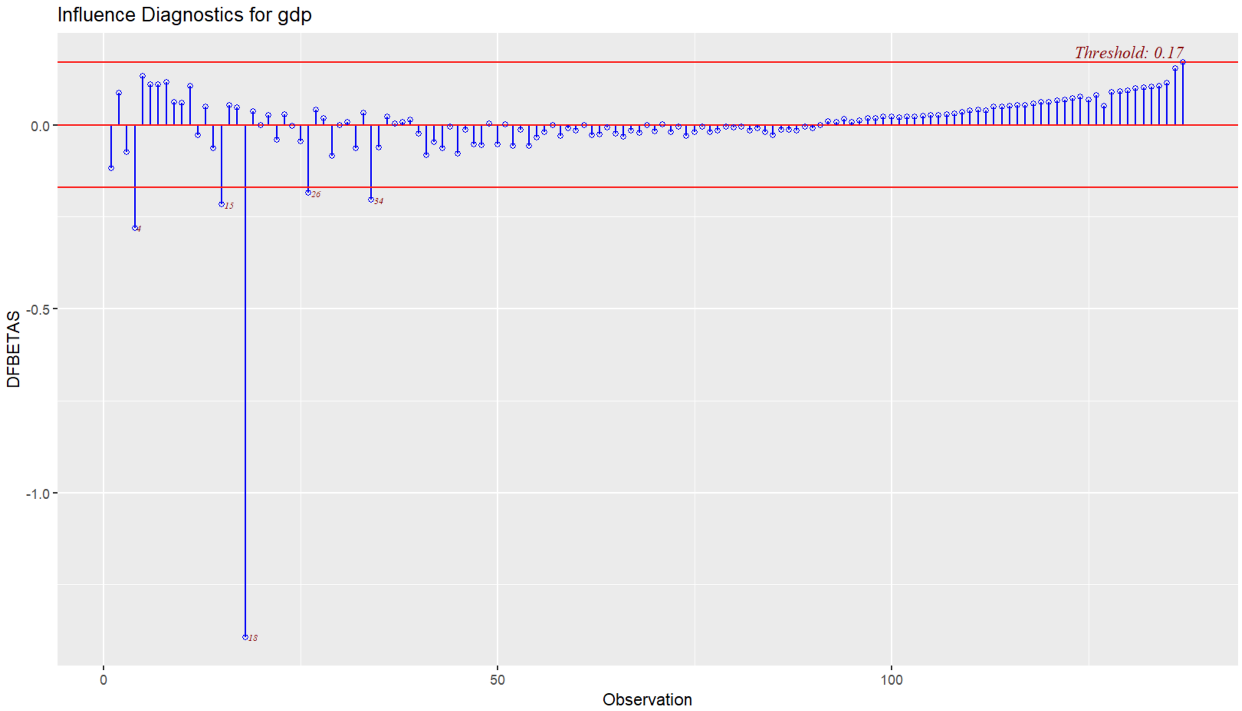
Linearity: the residuals vs fitted values plot shows a pattern, where the residuals start low at lower fitted values, go higher for middle fitted values, and then go back down for higher fitted values. The linearity assumption appears to be violated.

Normality: The normal Q-Q plot follows a straight line, it seems the normality assumption is met.

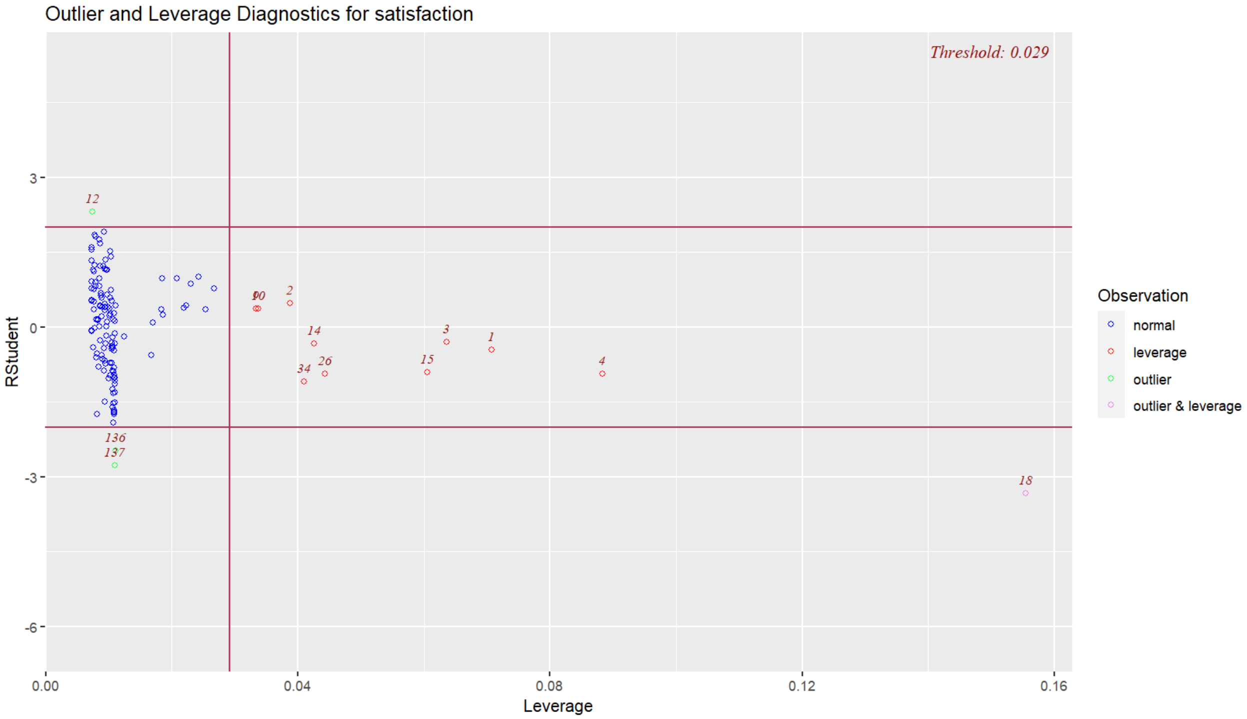
Homoscedasticity: There seems to be a funnel shape in the scale-location plot with more scatter for lower fitted values and less scatter for higher fitted values. This indicates presence of heteroscedasticity.

2f. [2 points] Assess whether there are observations that are influential. Which observations are these? Re-run your model from 2b without these influential observations and report how the regression equation changes.

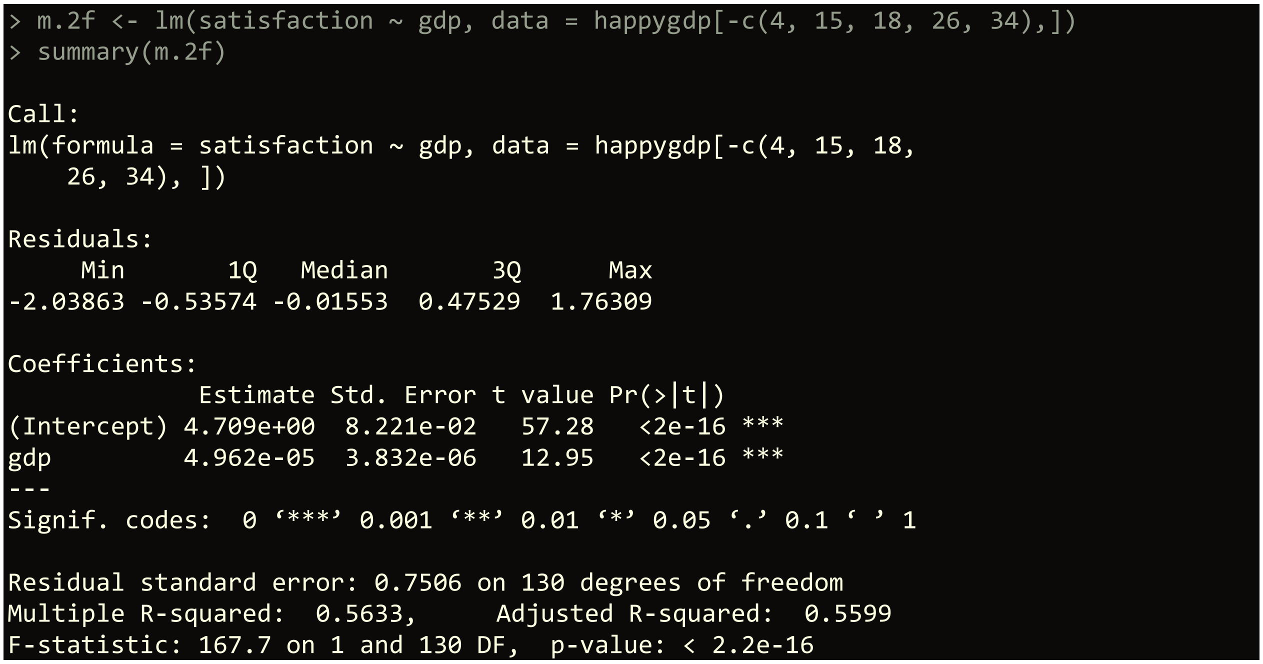
```
> ols_plot_dfbetas(m.2b, F)
```



```
> ols_plot_resid_lev(m.2b)
```



From the DFBETAS plot, it looks like 18 is highly influential, but also 4, 15, 26, and 34. From the leverage vs studentized residual plot, it confirms that observation 18 has both high leverage and is an outlier. I will remove 18, along with 4, 15, 26, and 34 and see how the regression line changes:



```
> m.2f <- lm(satisfaction ~ gdp, data = happygdp[-c(4, 15, 18, 26, 34),])
> summary(m.2f)

Call:
lm(formula = satisfaction ~ gdp, data = happygdp[-c(4, 15, 18,
    26, 34), ])

Residuals:
     Min       1Q   Median       3Q      Max
-2.03863 -0.53574 -0.01553  0.47529  1.76309

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.709e+00  8.221e-02   57.28   <2e-16 ***
gdp         4.962e-05  3.832e-06   12.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7506 on 130 degrees of freedom
Multiple R-squared:  0.5633,     Adjusted R-squared:  0.5599
F-statistic: 167.7 on 1 and 130 DF,  p-value: < 2.2e-16
```

Excluding the influential observations, the regression equation now becomes:

$$\hat{Y} = 4.71 + 0.00005X$$

The intercept decreased slightly from 4.793 to 4.709, and the slope increased slightly from 4.087e-05 to 4.962e-05. Overall the equation for the regression line did not change much.

2g. [1 point] Based on your responses to 2e and 2f, what do you think about the conclusions the authors reached?

Based on my responses to 2e and 2f, the assumptions of a linear regression are not met, namely linearity and homoscedasticity. With that in mind, I am hesitant to trust a linear regression model of happiness vs GDP as well as any conclusions the author drew from it. I think that to improve the model, the author could try to transform the GDP variable to make it linear, or include more predictor variables to get a better estimate of happiness. The author also hints that the trend holds differently for different wealth levels, perhaps the author could split the data by wealth levels, and then analyze and interpret the results for the categories separately.

Question 3 [11 points]

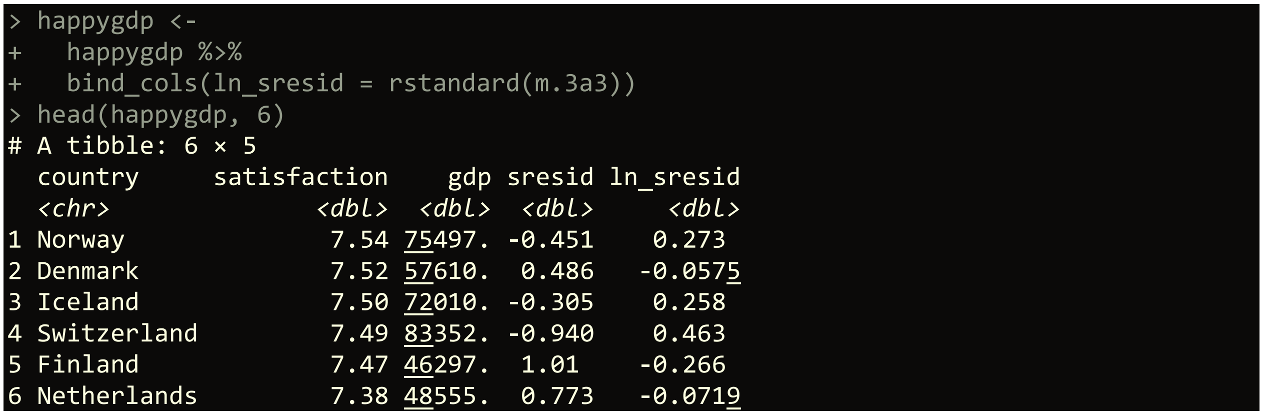
Try to create a better model than the one the authors created.

3a. [2 points] Report the R^2 of the following models:

- Your model in 2b
- A regression of satisfaction on transformed GDP (try 2 different transformations you think may work well)
- A regression of satisfaction on ln(GDP)

Transformation of SBP	R-squared
None	0.54
Square root	0.67
Tenth root	0.73
Natural log	0.73

3b. [1 point] Add the residuals from the model with ln(GDP) to your data set.

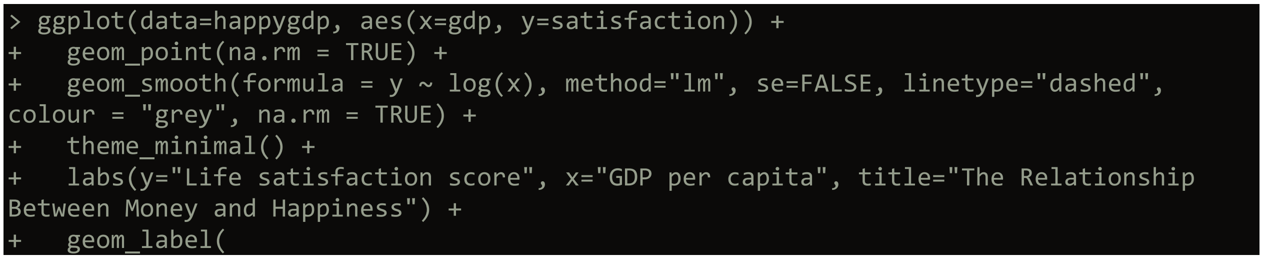


```
> happygdp <-
+   happygdp %>%
+   bind_cols(ln_sresid = rstandard(m.3a3))
> head(happygdp, 6)
# A tibble: 6 × 5
  country     satisfaction    gdp sresid ln_sresid
  <chr>              <dbl>  <dbl>  <dbl>     <dbl>
1 Norway              7.54 75497. -0.451    0.273
2 Denmark             7.52 57610.  0.486   -0.0575
3 Iceland             7.50 72010. -0.305    0.258
4 Switzerland         7.49 83352. -0.940    0.463
5 Finland             7.47 46297.  1.01    -0.266
6 Netherlands         7.38 48555.  0.773   -0.0719
```

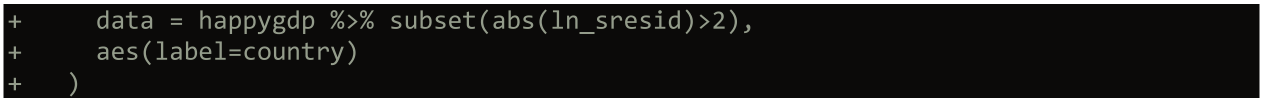
3c. [3 points] Update your plot as follows:

- Add the trendline from your regression model
  Hint: geom_smooth(method='lm', formula='y~log(x)')
- Add labels for all countries that have a standardized residual that is greater than 2 in magnitude.

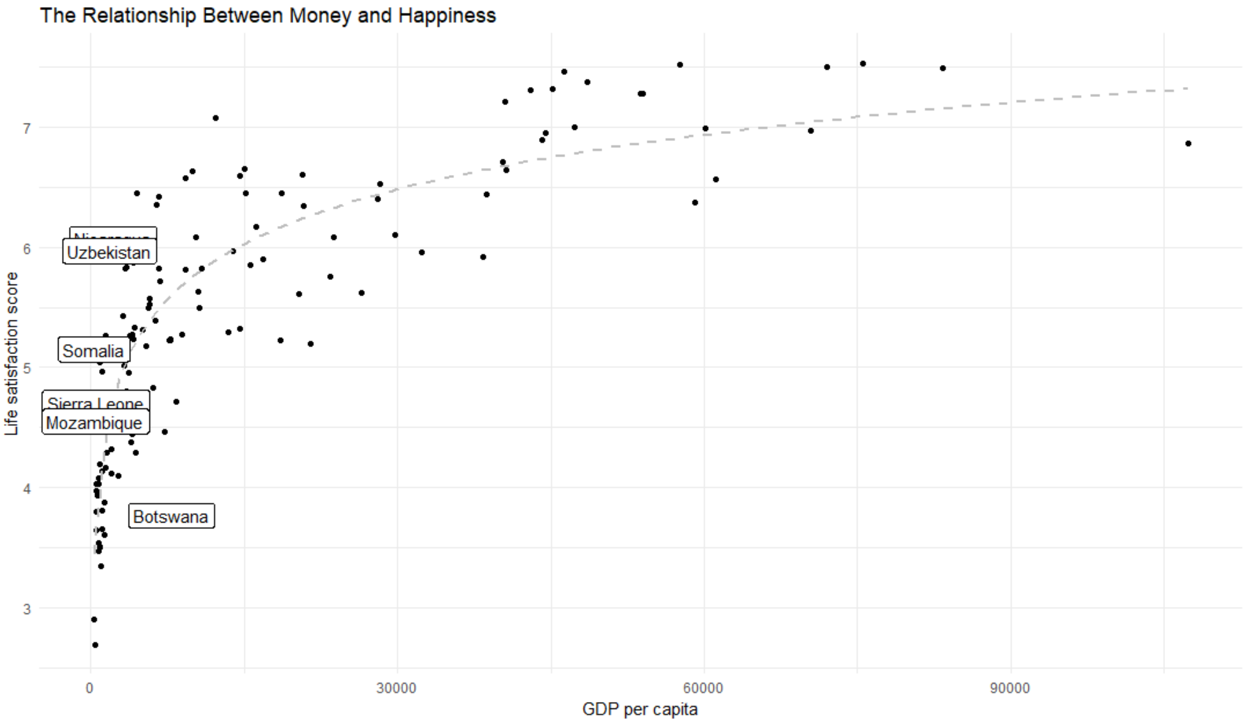
Compare the outliers in this model to the outliers from Question 2.



```
> ggplot(data=happygdp, aes(x=gdp, y=satisfaction)) +
+   geom_point(na.rm = TRUE) +
+   geom_smooth(formula = y ~ log(x), method="lm", se=FALSE, linetype="dashed", 
colour = "grey", na.rm = TRUE) +
+   theme_minimal() +
+   labs(y="Life satisfaction score", x="GDP per capita", title="The Relationship 
Between Money and Happiness") +
+   geom_label(
```



```
+     data = happygdp %>% subset(abs(ln_sresid)>2),
+     aes(label=country)
+   )
```

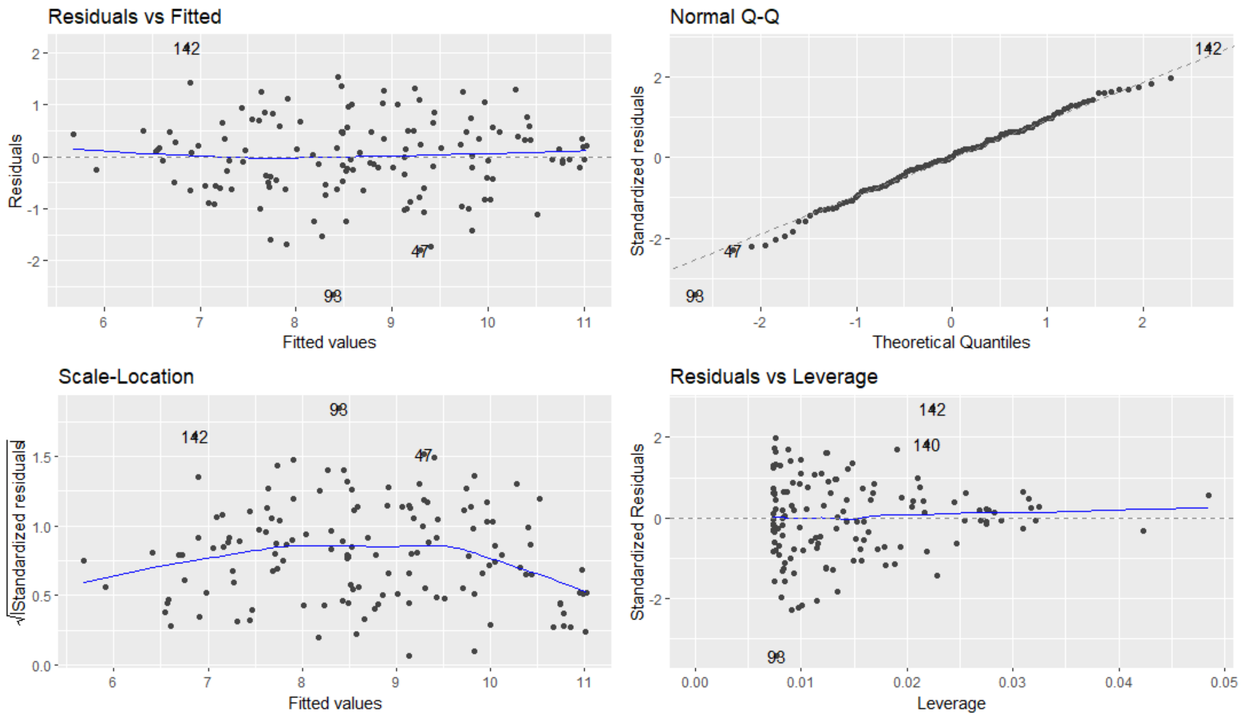


There are two more outliers when using residuals of the log transformed satisfaction versus with no transformation. None of the outliers in this model (Nicaragua, Uzbekistan, Somalia, Sierra Leone, Mozambique, and Botswana) overlap with the outliers in the non-transformed model (Costa Rica, Luxembourg, Burundi, and Central African Republic). Additionally, none of the residuals are from extremely wealthy countries, whereas before, Luxembourg was an outlier.

3d. [2 points] Describe (using plain language) the relationship between GDP and satisfaction. In Question 2, Luxembourg had the largest residual. Explain why you think Luxembourg is, or is not, an outlier with regard to the *true* relationship between satisfaction and GDP.

As a country's wealth increases by $1 per person, the country's overall satisfaction score is predicted to increase by 10.2%. Luxembourg is not an outlier in the true relationship between satisfaction and GDP because satisfaction is better defined as increasing as a function of the log of GDP. This means satisfaction increases more for changes in lower amounts of GDP per capita (e.g. $10K to $15K) and increases less as GDP per capita is higher (e.g. $90K to $95K).

3e. [3 points] Assess the assumptions of linear regression and report any influential countries.


Residuals vs Fitted

Normal Q-Q

Scale-Location

Residuals vs Leverage

Linearity: the residuals vs fitted plot show a random scatter around a mean line of 0, the assumption of linearity is met.

Normality: the normal Q-Q plot shows the points following the straight line very well, the assumption of normality is met.

Homoscedasticity: the scale-location plot shows no pattern overall, and generally looks like a random scatter; the assumption of homoscedasticity is met.