Question 1)

Analysis of figure 3.

Since the analysis of the assumptions are based on the figures alone, it is difficult to discern if the normality of residuals or independence of observations assumptions were met so only linearity and homoscedasticity assumptions were inspected.

The first regression analysis (VIIRS vs GDP) is linear and the variance around the regression line appears to be constant throughout, thus regression is suitable.

Both the second regression analysis (VIIRS vs Population) and the fourth analysis (VIIRS vs CO_2) seems to curve upwards slightly but this cannot be confirmed with certainty, therefore linearity will be assumed since there is no obvious curve. However, in terms of homoscedasticity, as the value of x increases, the variance of residuals around the regression line for both regression analysis also increases thus violating the homoscedasticity assumption.

The third regression analysis (VIIRS vs Road) is linear but as the value of *x* (Road) increases, so does the variance resulting in a violation of the homoscedasticity assumption.

Overall, from inspecting figure 3, only the first regression analysis is suitable for linear regression.

Analysis of figure 11.

From figure 11, it can be noted that all the regression lines above are linear. Although there are some data points that could imply a slight curve in LOG (VIIRS) vs LOG (GDP), the data is too sparse in that region, and we cannot confirm the violation through other analytical methods thus linearity is assumed.

In terms of homoscedasticity, it is uncertain if this assumption is breached in the LOG(VIIRS) vs LOG(CO_2 density) as it appears that the variance of data points start to expand towards the middle of the regression line but this cannot be confirmed thoroughly so homoscedasticity will be assumed.

The only regression analysis that is obviously violating the homoscedastic assumption is the third regression analysis (LOG VIRRS vs LOG Road) where the data points start spread out and converge inwards, resulting in an uneven variance of residuals and violating the homoscedastic assumption.

Question 2)

After log transforming both outcome variable (GDP) and covariate and comparing the result with the original regression. It is believed that the log transformed regression is more suitable for the data, as seen in fig 1. a better linear relationship can be observed between outcome and covariate after transformation where the data points are better spread out.

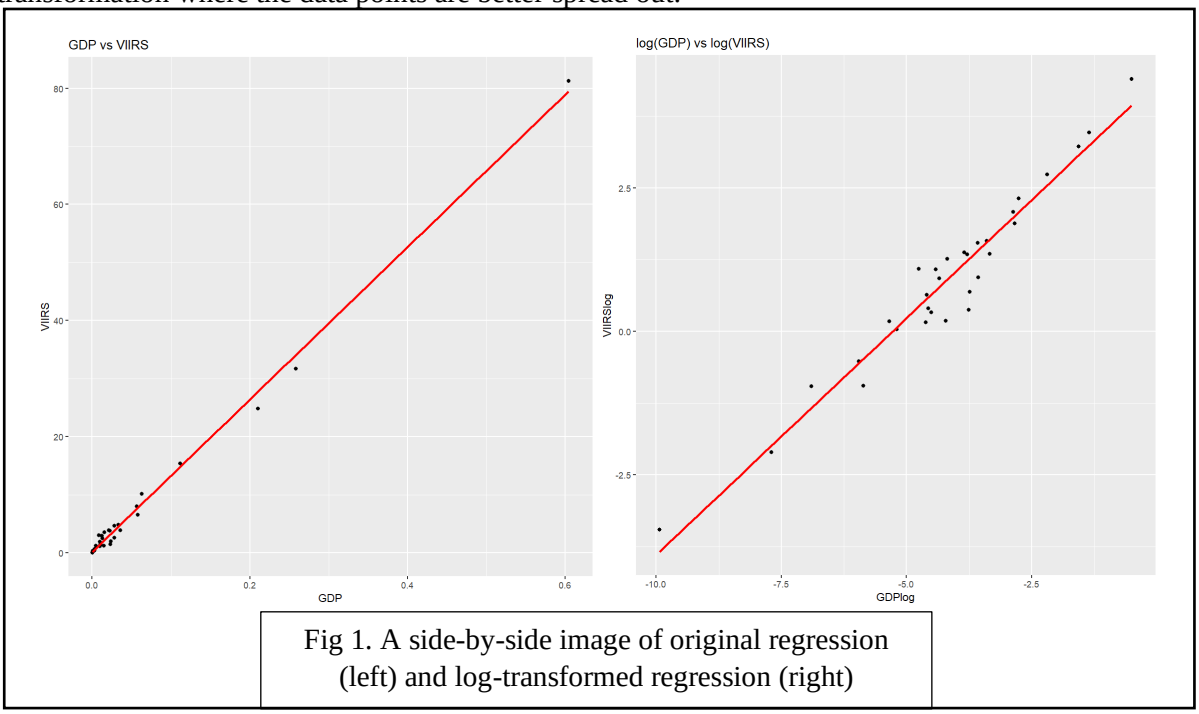

Fig 1. A side-by-side image of original regression (left) and log-transformed regression (right)

Additionally, after log transforming the original regression, it could be noted that the more extreme data values had both their residual values and leverage greatly reduced (Figure 2), which means less influence on the parameters and a more accurate model.

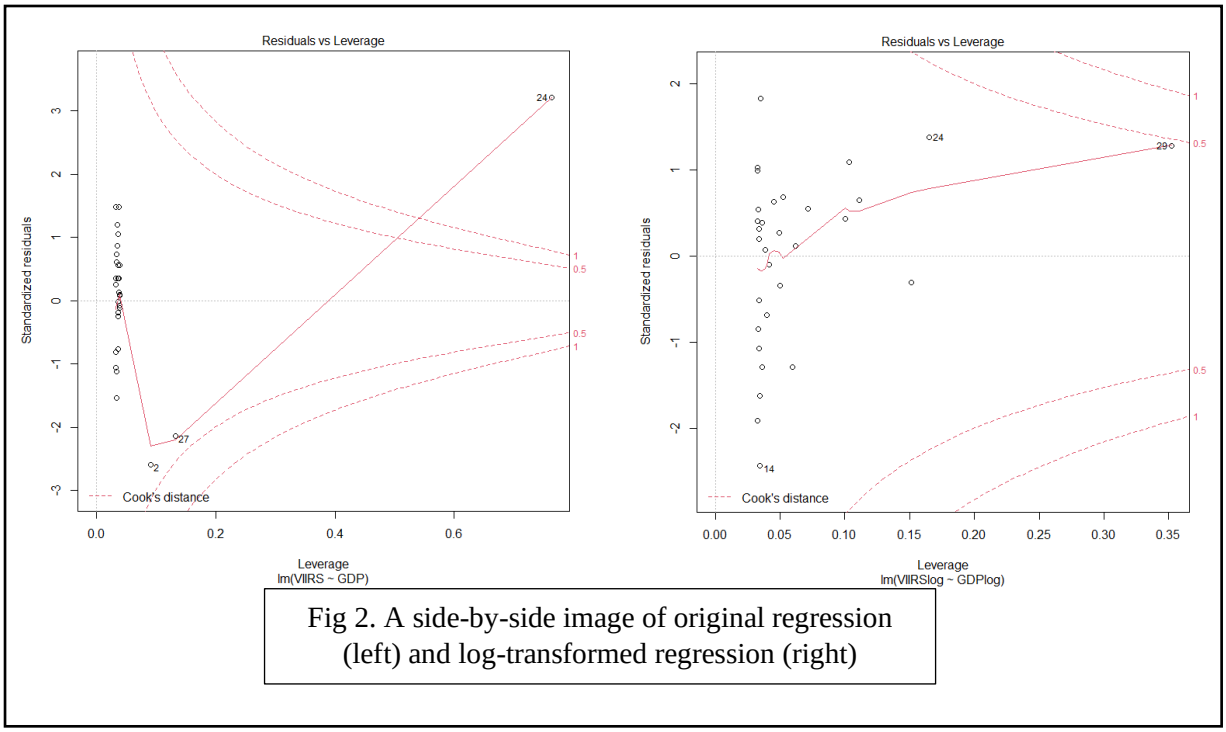

Fig 2. A side-by-side image of original regression (left) and log-transformed regression (right)

Conducting a summary of both also reveals that even with a minor reduction of R^2 value after transformation, the residual standard error has been largely reduced after the log-log transformation showing evidence that the transformed regression is a better predictor and fits the dataset better than the original regression.

Table 1. Summary of the parameters from both regressions

	β_0	β_1 (Lwr CI – Upr CI)	β_1 std. error	R^2	R^2 std. error	t value	p value
Original Regression	0.1511	131.26 (127 – 135)	1.8257	0.9944	1.173	71.902	<0.001
Log Transformed Regression	4.34920	0.82415 (0.75-0.90)	0.03631	0.9467	0.3681	22.7	<0.001

After choosing the log-log regression model, a prediction of VIIRS density using a GDP of log (0.01) \$B/km^2 was calculated using R, producing the following results.

	Mean	Lower Interval	Upper Interval
Confidence intervals	0.5538547	0.4153231	0.6923864
Prediction intervals	0.5538547	-0.2116542	1.319364

These results were then back transformed using the following formula,

$$\log(y) = 4.34 + 0.82415 \log(x)$$

$$\text{therefore}, y = \exp[4.34 + 0.82415 \log(x)]$$

	Mean	Lower Interval	Upper Interval
Confidence intervals	**1.74**	**1.51**	**2.00**
Prediction intervals	**1.74**	**0.81**	**3.74**

Question 3)

To summarise the data presented above, using the transformed regression model, it can be stated that there is strong evidence that VIIRS density increases as GDP increases ($p < 0.001$) and that for every 1% increase of GDP there will be an expected 0.82% increase in VIIRS density. We are 95% confident that the true value of this expected increase being between 0.75% to 0.90%. From the R^2 value observed, it can also be noted that GDP accounts for about 95% variability that occurs in VIIRS density. Using the regression model created, a prediction was made that the expected VIIRS density for a GDP of 0.01\$B/Km^2 will result in an expected VIIRS density of 1.74, with this true value being somewhere between 1.51 and 2. 95% of any individual data value with a GDP of 0.01 could have a predicted VIIRS density between 0.81 and 3.74.

Question 4)

To meet the criteria for confounding the relationship between light density and GDP, CO2 density would have to be both associated with the outcome variable (light density) and the covariate (GDP) and must not lie on the causal pathway between GDP and Light density.

Running a regression with both GDP and CO2 density as explanatory variables (both log transformed) and log transformed light density as an outcome variable we get the following results,

	β_0	β_1 (Lwr CI – Upr CI)	β_1 std. error	R^2	R^2 std. error	t value	p value
GDPlog	9.89	0.5860 (0.342 – 0.83)	0.1189	0.9539	0.3484	4.927	**<0.001*** **
CO2log	9.89	0.3052 (0.006-0.6042)	0.1460	0.9539	0.3484	2.091	**0.0457***

While GDP still has a statistically significant relationship with light density, when adjusting for CO2 density. The previously expected 0.82% increase was reduced to a 0.59% increase per 1% increase of GDP instead. CO2 density is also shown to have evidence of an association between itself and light density with a p-value of 0.0457 and causes an expected increase of 0.31% in light density per 1% increase in CO2 density. Overall, CO2 density has an association with light density and overtly affects the association between GDP and light density as well. Therefore, to obtain an unbiased estimate of the relationship between light density and GDP, CO2 density should be adjusted for.

Question 5)

Assuming $\beta_2 = 2\beta_1$, manipulating the equation for the multiple linear regression model with the assumption in mind provides the following constrained regression model.

$$y = \beta_0 + \beta_1(x_1) + \beta_2(x_2)$$

$$y = \beta_0 + \beta_1(x_1) + 2\beta_1(x_2)$$

$$y = \beta_0 + \beta_1(x_1) + 2\beta_1(x_2)$$

$$y = \beta_0 + \beta_1[(x_1) + 2(x_2)]$$

$$\beta_0 + \beta_1 = \frac{y}{(x_1) + 2(x_2)}$$

With the following matrix parameters.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & 2(x_{21}) \\ 1 & x_{12} & 2(x_{22}) \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & 2(x_{2n}) \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

APPENDIX

```
Assignment <- read.csv("C:/Users/Ka/Desktop/Stats Datasets/Regression/Assignment
dataset.csv")


library(ggplot2)

library(magrittr)


#Variables without modification

lm.VIIRSGDP <- lm(VIIRS ~ GDP, data = Assignment)


summary(lm.VIIRSGDP)

confint(lm.VIIRSGDP)

#B0 = 0.1511, B1 = 131.268

#R2 = 0.9944, GDP accounts for 99% variability in VIIRS

#There is a 95% chance that the real value of B1 lies between 127.53



plot(Assignment$GDP,Assignment$VIIRS)

plot(VIIRS ~ GDP, data = Assignment)


#Shows linearity

ggplot(Assignment, aes(x = GDP, y = VIIRS)) +

  geom_jitter() +

  labs(title = "GDP vs VIIRS") +

    geom_smooth(method = "lm", color = "red",   se = FALSE)


plot(lm.VIIRSGDP, 1)

plot(lm.VIIRSGDP, 2)

plot(lm.VIIRSGDP, 5)

hist(lm.VIIRSGDP$residuals, breaks = 10)



#-------------------------------------------------------------------------

#log variables VIIRS and GDP

Assignment$VIIRSlog <- log(Assignment$VIIRS)

Assignment$GDPlog <- log(Assignment$GDP)
```

```
lm.VIIRSGDPlog <- lm(VIIRSlog ~ GDPlog, data = Assignment)
summary(lm.VIIRSGDPlog)
confint(lm.VIIRSGDPlog)

#B0 = 4.349, B1 = 0.82415
#0.824% increase in VIIRS for every 1% increase in GDP

#R2 = 0.9467

#Shows linearity
ggplot(Assignment, aes(x = GDPlog, y = VIIRSlog)) +
  geom_jitter() +
  labs(title = "log(GDP) vs log(VIIRS)") +
  geom_smooth(method = "lm", color = "red",   se = FALSE)

summary(lm.VIIRSGDPlog)
plot(lm.VIIRSGDPlog, 1)
plot(lm.VIIRSGDPlog, 2)
plot(lm.VIIRSGDPlog, 5)
hist(lm.VIIRSGDPlog$residuals, breaks = 10)


#Prediction of 0.01 $B/km GDP
Prediction <- data.frame(GDPlog=log(0.01))
predict(lm.VIIRSGDPlog, Prediction, interval="confidence")
predict(lm.VIIRSGDPlog, Prediction, interval="prediction")

#------------------------------------------------------------------------------

#Q4
plot(Assignment$CO2, Assignment$VIIRS)
plot(Assignment$CO2, Assignment$VIIRSlog)
Assignment$CO2log <- log(Assignment$CO2)
plot(Assignment$CO2log, Assignment$VIIRSlog)
plot(Assignment$GDPlog, Assignment$CO2log)

ggplot(data=Assignment) +
```

```
  geom_smooth(mapping = aes(x = GDPlog, y= CO2log))

lm.Multiple <- lm(VIIRSlog ~ GDPlog + CO2log, data = Assignment)
summary(lm.Multiple)
confint(lm.Multiple)
#-------------------------------------------------------------------------------

#Q5 experiment [not successful]

summary(lm(VIIRS ~ GDP + CO2, data = Assignment))

Assignment$CO22 <- Assignment$CO2log*2

summary(lm(VIIRSlog ~ GDPlog + CO22, data = Assignment))
```