Economic Influence on Fertility: A Linear Regression Analysis

Aqsa Noreen

2023-10-20

Data visualization and Pre-Processing

```
library(ggplot2)

# Read the data
data <- read.table("UN.txt", header = TRUE)
head(data)
```

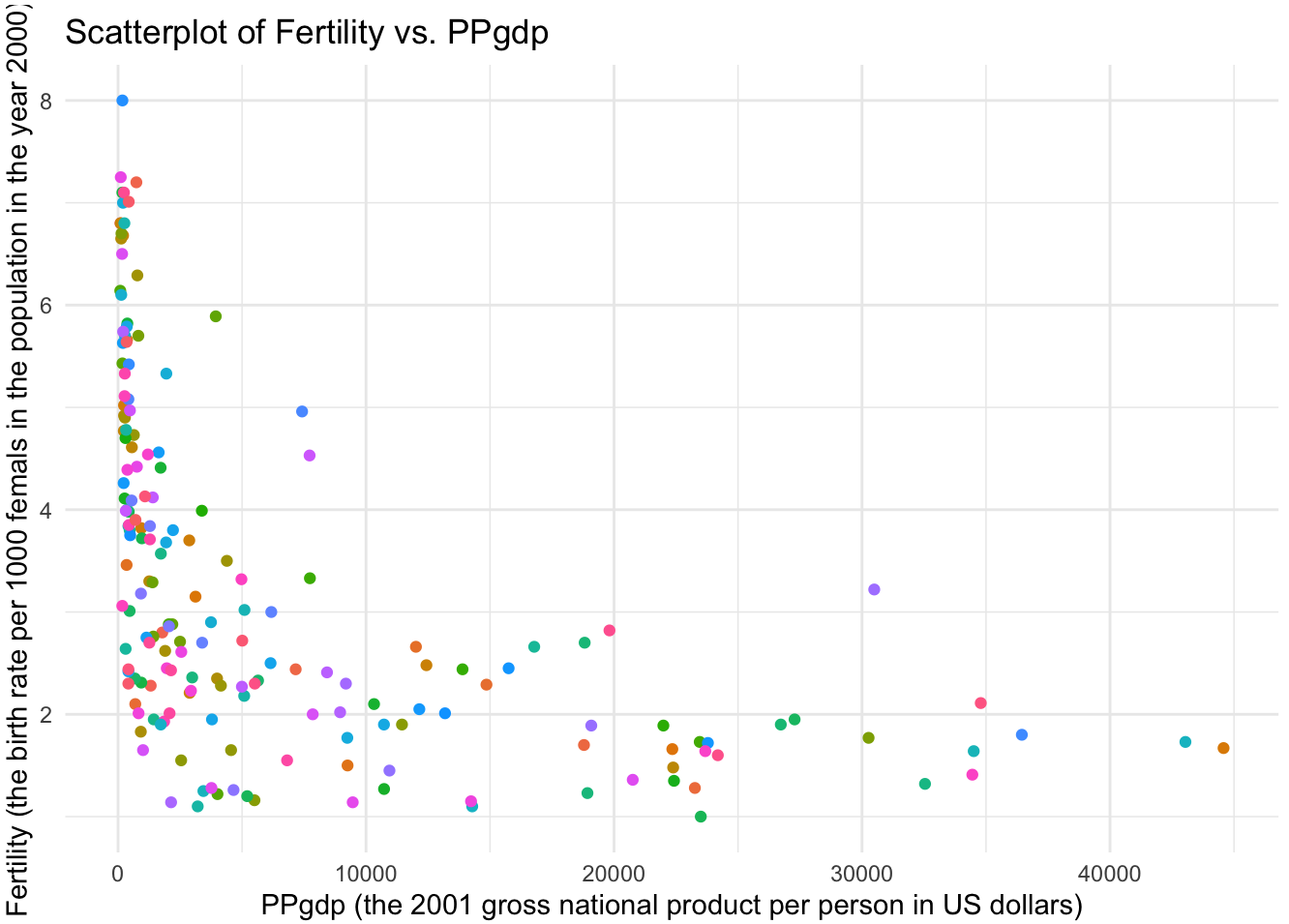
	Locality <chr>	Fertility <dbl>	PPgdp <int>
1	Afghanistan	6.80	98
2	Albania	2.28	1317
3	Algeria	2.80	1784
4	Angola	7.20	739
5	Argentina	2.44	7163
6	Australia	1.70	18788

6 rows

```
apply(data, 2, function(x) sum(is.na(x)))
```

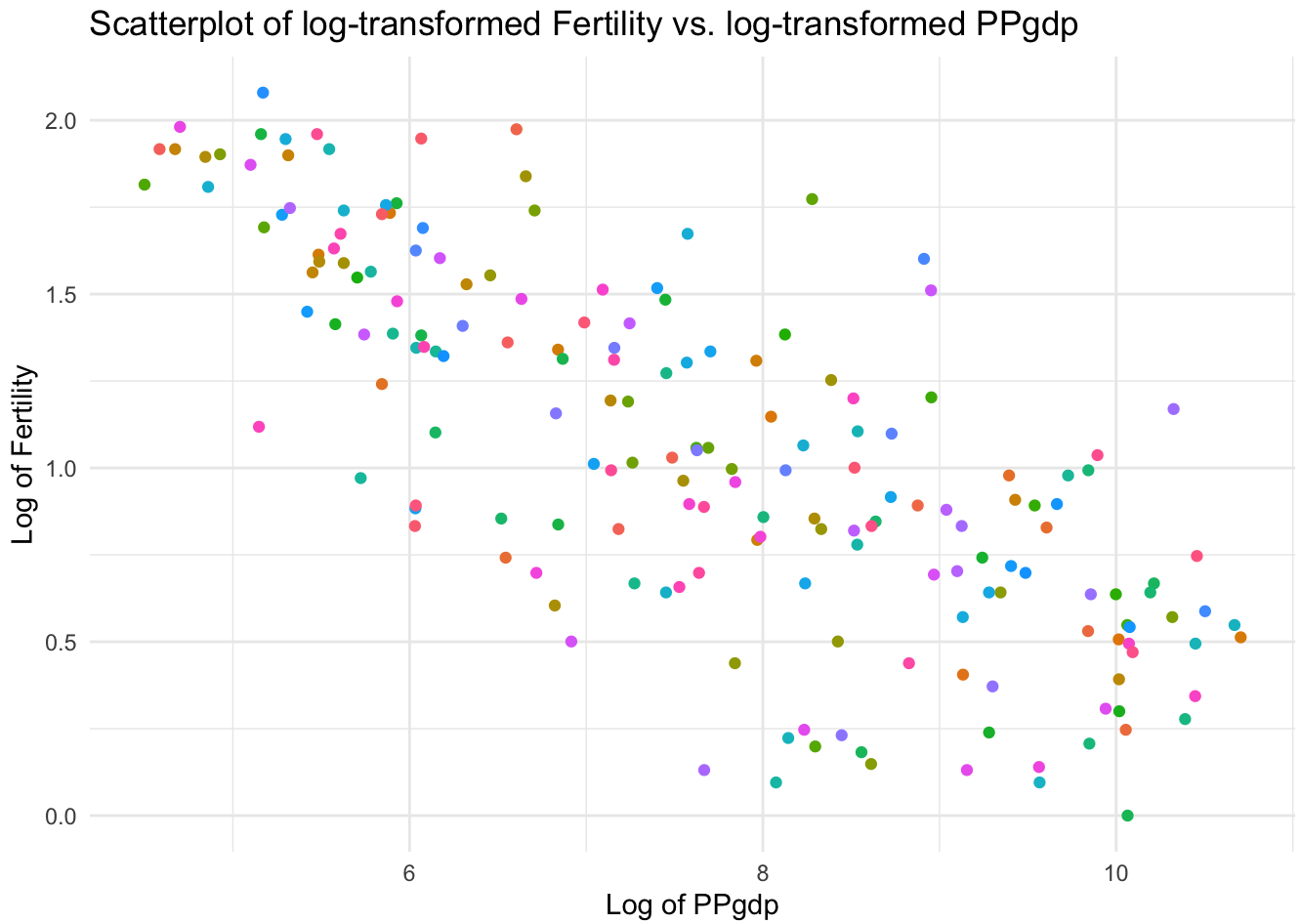
```
##  Locality Fertility     PPgdp
##         0         0         0
```

```
#Draw the scatterplot of Fertility on the vertical axis versus PPgdp on the horizontal a
xis
X=ggplot(data, aes(x = PPgdp, y = Fertility)) +
  geom_point(aes(color = Locality)) +
  labs(title = "Scatterplot of Fertility vs. PPgdp",
       x = "PPgdp (the 2001 gross national product per person in US dollars)",
       y = "Fertility (the birth rate per 1000 femals in the population in the year 200
0)") +
  theme_minimal() +
  theme(legend.position = "none")  # Hiding the legend since there are toooo many countr
ies (184 localities)
print(X)
```



We can see a general downward trend which displays a negative correlation between PPgdp and Frtility since as PPgdp increases, Fertility tends to decrease.We can see that initially when PPgdp is at its lower values the Fertility rates are more spread out and range from low to very high. But at the PPgdp increases the fertility rates converge and they cluster closer to the lower end of the scale now. This plot illustrate that countries with higher income per person tend to have lower fertility rates. There are some outliers such as countries with very high PPgdp and very low fertility, along with low PPgdp and wither very high or low fertility (Can potentially influence the fit of our regression model) Relationship between PPgdp and Fertility does not seem to be strictly linear (mainly at lower values of PPgdp). Although we see a clear negative relation between Fertility and PPgdp, I think that a simple linear model might not be the best fit for this data due to the apparent non-linearity (especially at lower values of PPgdp). We will need to apply a Logarithmic transformation to Fertility and PPgdp to get the underlying relationship.

```
# Transformation
data$log_Fertility <- log(data$Fertility)
data$log_PPgdp <- log(data$PPgdp)

# scatter plot of the transformed variables
X=ggplot(data, aes(x = log_PPgdp, y = log_Fertility)) +
  geom_point(aes(color = Locality)) +
  labs(title = "Scatterplot of log-transformed Fertility vs. log-transformed PPgdp",
       x = "Log of PPgdp",
       y = "Log of Fertility") +
  theme_minimal() +
  theme(legend.position = "none")
print(X)
```



Our main goal of this transformation is to help linearize the relationship between Fertility and PPgdp (realtion b/w predictor variable and response variable). And make our data more suitable for the linear regression model, log transformation help us do that. We also see a potential heteroscedasticity in our original scatterplot where the spread of data points around any potential regression line does not seem to be consistent. As mentioned earlier that for the low value of PPgdp the points are tightly clustered but as PPgdp increases the dispersal of the Fertility values seem yo decrease. We can see that in the Log transformed scatterplot, the spread of the data points appears more even across different values of PPgdp log. Homoscedasticity is achieved since now we have more consistent spread. Log transformation also helped reduce the impact of extreme outliers that can have an impact on the regression line. Comments on the Plot: Now the relationship between to variables appear more linear although we do still see some scatter, despite that the downward trend is more clear and has a consistent trajectory across PPgdp. Also the variation of Fertility across PPgdp is more uniform now as well. There is now only a moderate influence of the outliers.

Model Fitting and Diagnostics

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

```
# Mean
x_bar <- mean(data$log_PPgdp)
y_bar <- mean(data$log_Fertility)

#numerator and denominator for beta_1
#num <- sum((data$log_Fertility - x_bar)*(data$log_PPgdp - y_bar))

#denom <- sum((data$log_Fertility - x_bar)^2)

num <- sum((data$log_PPgdp - x_bar) * (data$log_Fertility - y_bar))
denom <- sum((data$log_PPgdp - x_bar)^2)

# beta_1 and beta_0
beta_1 <- num/denom
beta_0 <- y_bar - beta_1 * x_bar

# prediction
predict_y <- beta_0 + beta_1 * data$log_PPgdp

#R-squared
SSR <- sum((data$log_Fertility - predict_y)^2)
SST <- sum((data$log_Fertility - y_bar)^2)
r_squar <- 1 - (SSR/SST)

cat("Slope:", beta_1, "\n")
```

```
## Slope: -0.2374852
```

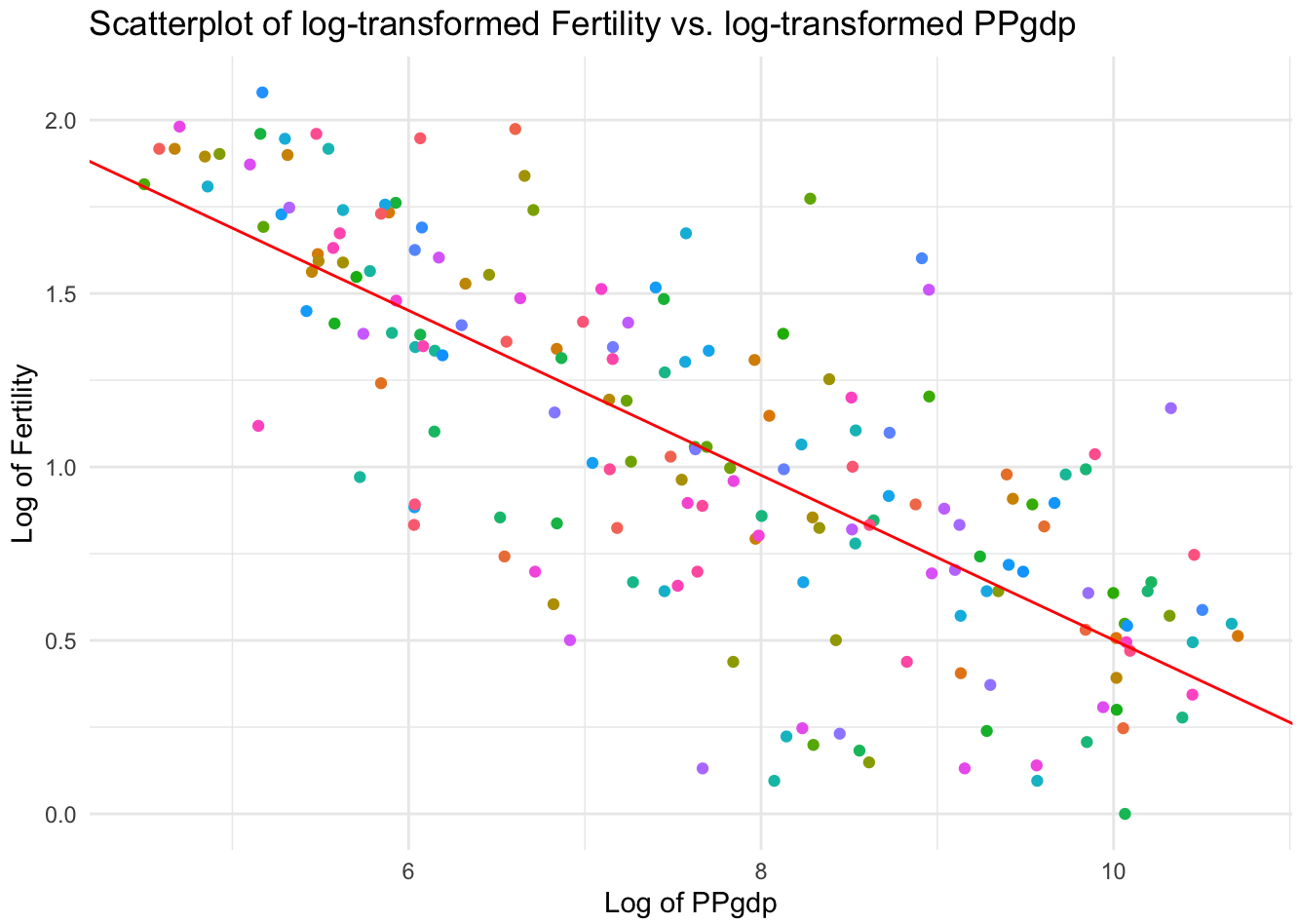
```
cat("Intercept:", beta_0, "\n")
```

```
## Intercept: 2.876072
```

```
cat("R-Squared:", r_squar, "\n")
```

```
## R-Squared: 0.5813292
```

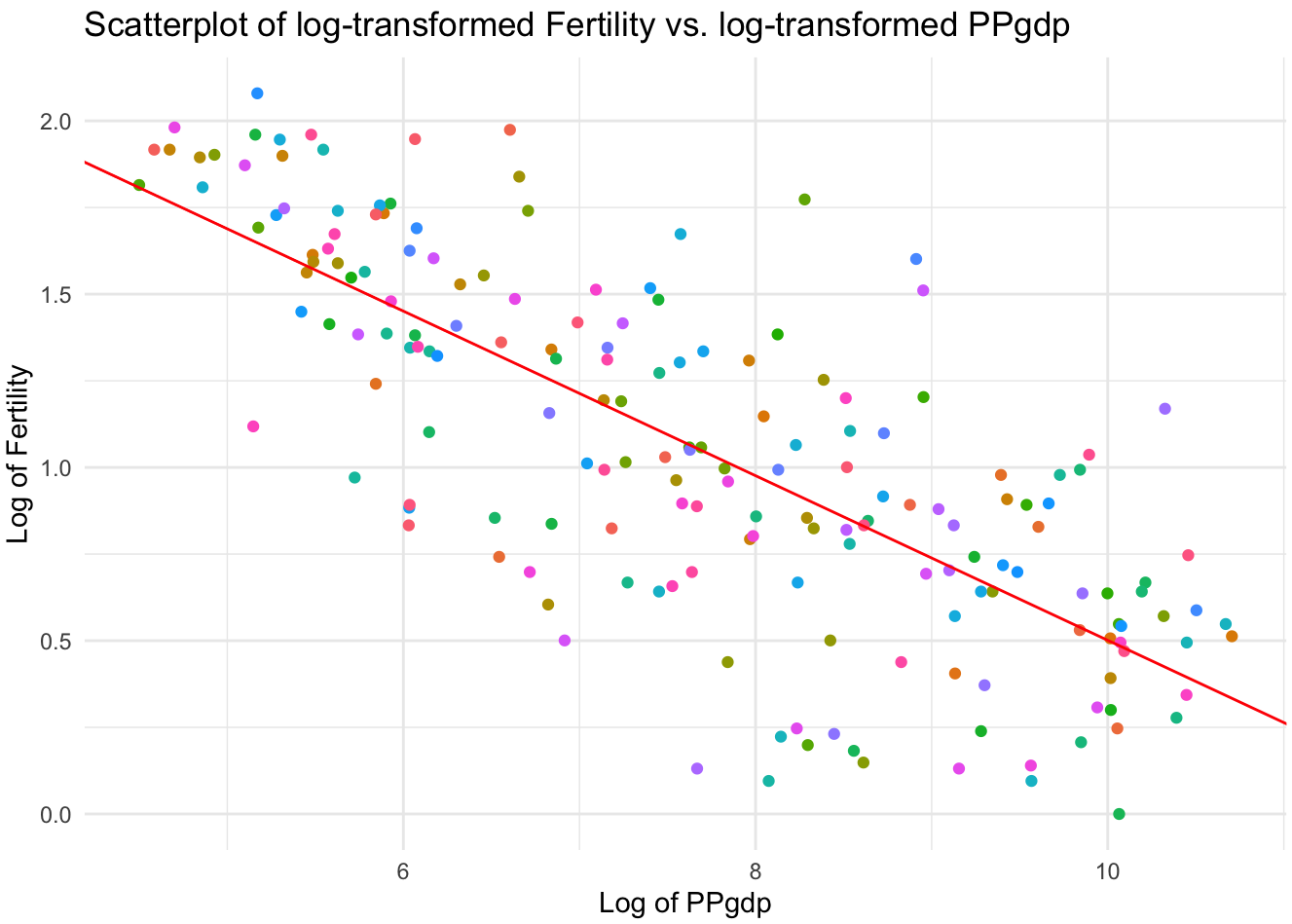
```
X + geom_abline(intercept = beta_0, slope = beta_1, color = "red")
```




```
#fit model using lm function
linear_model <- lm(log_Fertility ~ log_PPgdp, data=data)
print(summary(linear_model))
```

```
##
## Call:
## lm(formula = log_Fertility ~ log_PPgdp, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92398 -0.16996  0.03671  0.20633  0.86331
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.87607    0.11715   24.55   <2e-16 ***
## log_PPgdp   -0.23749    0.01494  -15.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3377 on 182 degrees of freedom
## Multiple R-squared:  0.5813, Adjusted R-squared:  0.579
## F-statistic: 252.7 on 1 and 182 DF,  p-value: < 2.2e-16
```

```
X + geom_abline(intercept = coef(linear_model)[1], slope = coef(linear_model)[2], color
= "red")
```



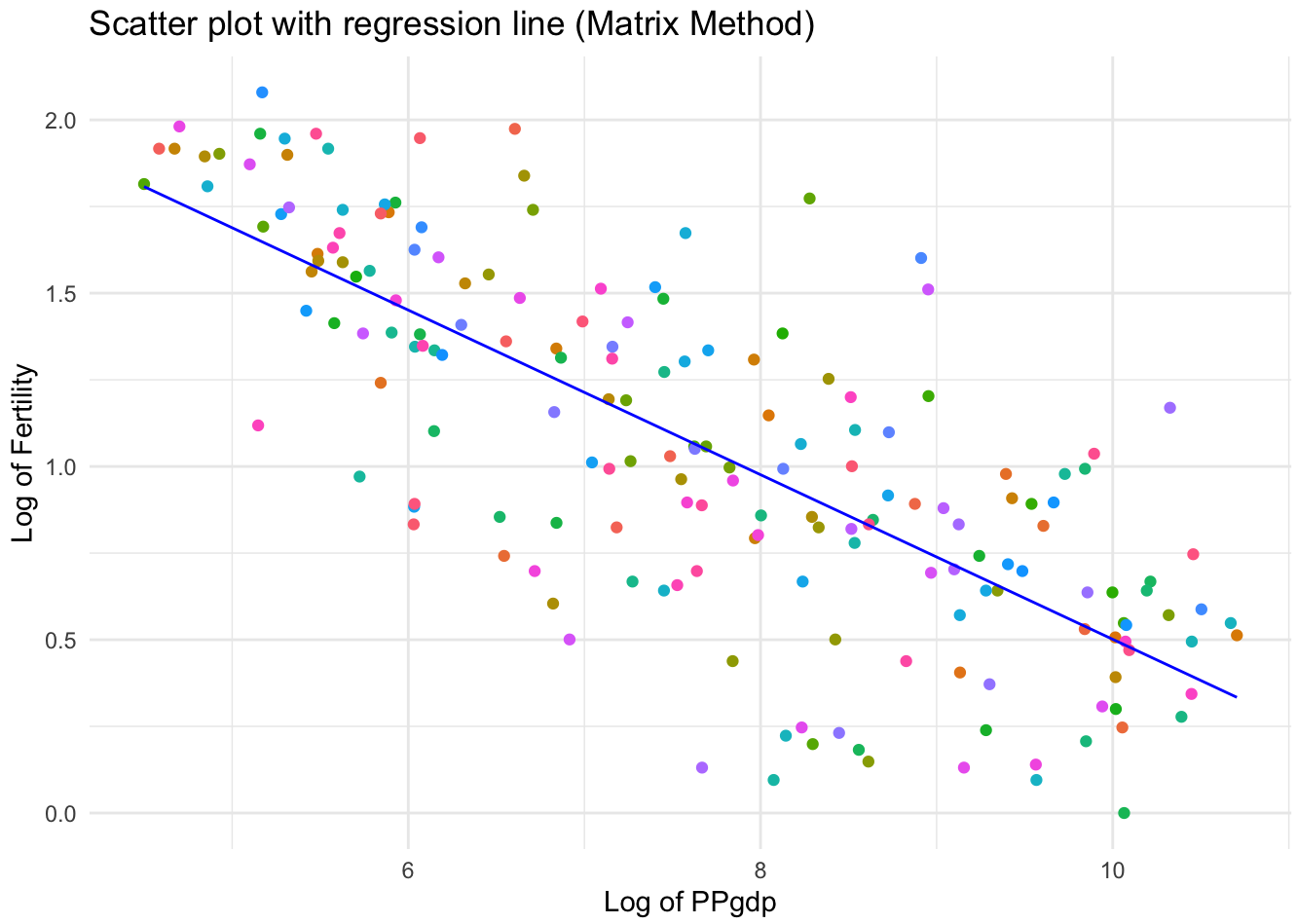

```
#Through matrix manipulation
X_matrix <- cbind(1, data$log_PPgdp)

Y_matrix <- data$log_Fertility

beta <- solve(t(X_matrix) %*% X_matrix) %*% t(X_matrix) %*% Y_matrix

cat("Slope (Matrix Method):", beta[2], "\n")
```

```
## Slope (Matrix Method): -0.2374852
```

```
cat("Intercept (Matrix Method):", beta[1], "\n")
```

```
## Intercept (Matrix Method): 2.876072
```

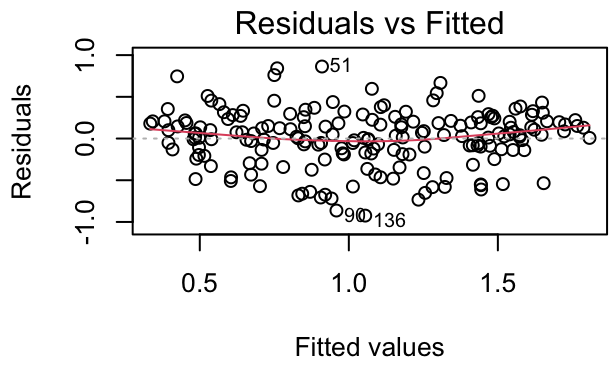
```
predicted_Y = X_matrix %*% beta

data$predicted_Y_matrix_method = as.vector(predicted_Y)

# Create a new plot with the updated data dataframe
X_new <- ggplot(data, aes(x = log_PPgdp, y = log_Fertility)) +
  geom_point(aes(color = Locality)) +
  labs(title = "Scatterplot of log-transformed Fertility vs. log-transformed PPgdp",
       x = "Log of PPgdp",
       y = "Log of Fertility") +
  theme_minimal() +
  theme(legend.position = "none")

# regression line matrix method
X_updated = X_new +
  geom_line(aes(x=log_PPgdp, y=predicted_Y_matrix_method), color="blue") +
  labs(title="Scatter plot with regression line (Matrix Method)")

print(X_updated)
```

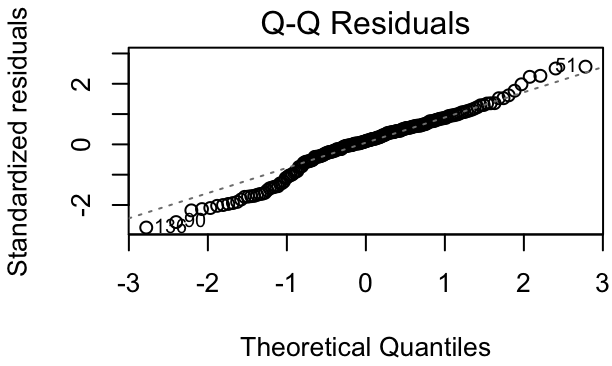


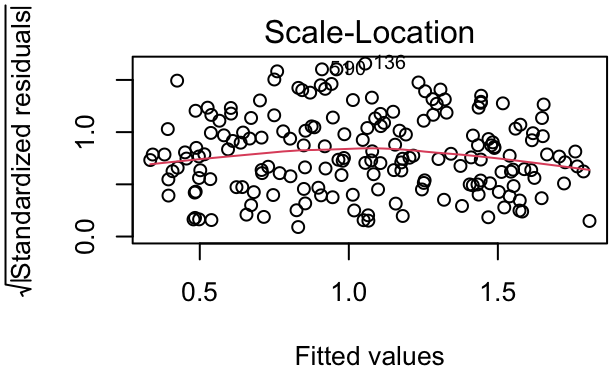
The regression lines for all three methods seem to provide a good and reasonable fit to the data. We can see a downward trend that indicate that as PPdgp increase the fertility tends to decrease. The spread of the residual appears to be fairly consistent.

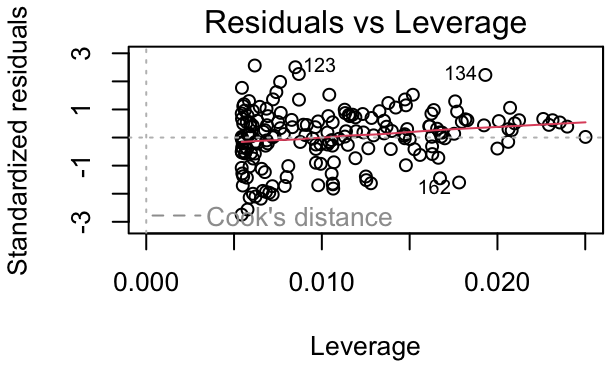
Draw the diagnostic plots and comment.

```
#plot(linear_model, which=1)
#plot(linear_model, which=2)

#plot(linear_model, which=3)
#plot(linear_model, which=4)

par(mfrow = c(2, 2))
plot(linear_model)
```










```
# Reference: http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/
```

Residual Vs Fitted: The residual is scattered around the lines without any distinct pattern indicating a linear relationship. Assumption of equal variance seems to be met. For the Normal QQ Plot, the points closely follow the straight line, but with slight deviations at the tails. Meaning that the residuals are approximately normally distributed with a slight concern about the normality in the tail. In the Scale-Location Plot there is not any distinct pattern meaning that the variences of the residual are approximately constant. For residuals vs leverage most of the data points are within the Cook's distance lines, suggesting there are no highly influential points.

Inference

Test whether there is a linear relationship between the transformed variables.

```
model_summary <- summary(linear_model)

model_summary$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  2.8760719 0.11714789  24.55078 1.186611e-59
## log_PPgdp   -0.2374852 0.01493916 -15.89683 3.004871e-36
```

Intercept: The estimated intercept is 2.8760719. This is the expected value of log_Fertility when log_PPgdp is 0. The standard error of the intercept is 0.11714789. The t-value for the intercept is 24.55078. The p-value for the intercept is 1.186611e-59 which is extremely close to 0. Given that this p-value is far below a common significance level, it is indicated that the intcpt is signficantly differennt from 0

Slope: The estimated slope is -0.2374852. This indicates that on average, for a unit increase in log_PPgdp, the log_Fertility decreases by approximately 0.2375. The standard error of the slope is 0.01493916. The t-value for the slope is -15.89683. The negative sign indicates that the estimated slope is below 0, which is in line with the negative estimate. The p-value for the slope is 3.004871e-36. This extremely low p-value means that the slope is significantly different from 0.

So overall we can say that, the negative slope suggests that there's a negative linear relationship between log_PPgdp and log_Fertility. As log_PPgdp increases, log_Fertility tends to decrease.

Given the extremely low p-value for log_PPgdp (much less than 0.05), we DO REJECT the null hypothesis. This means that there is a statistically significant linear relationship between the transformed variables log_Fertility and log_PPgdp.

Provide a 99% confidence interval on the expected Fertility for a region with PPgdp 20,000 US dollars in 2001.

```
log_ppgdp_20000 <- log(20000)
confidence_interval <- predict(linear_model, newdata = data.frame(log_PPgdp = log_ppgdp_
20000),
                               interval = "confidence", level = 0.99)

# transfer back to the original scale
fertility_lower_bound <- exp(confidence_interval[1, "lwr"])
fertility_upper_bound <- exp(confidence_interval[1, "upr"])

cat("fertility_lower_bound:", fertility_lower_bound, "\n")
```

```
## fertility_lower_bound: 1.515193
```

```
cat("fertility_upper_bound:", fertility_upper_bound, "\n")
```

```
## fertility_upper_bound: 1.882757
```

We are 99% confident that the true expected Fertility rate in terms of births per 1000 females in the year 2000 for a region with a PPgdp of 20,000 US dollars in 2001 lies between approximately 1.515193 and 1.882757.

Provide a 95% confidence band for the relation between the expected Fertility and PPgdp. Add the bands to the scatter plot of the original data.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

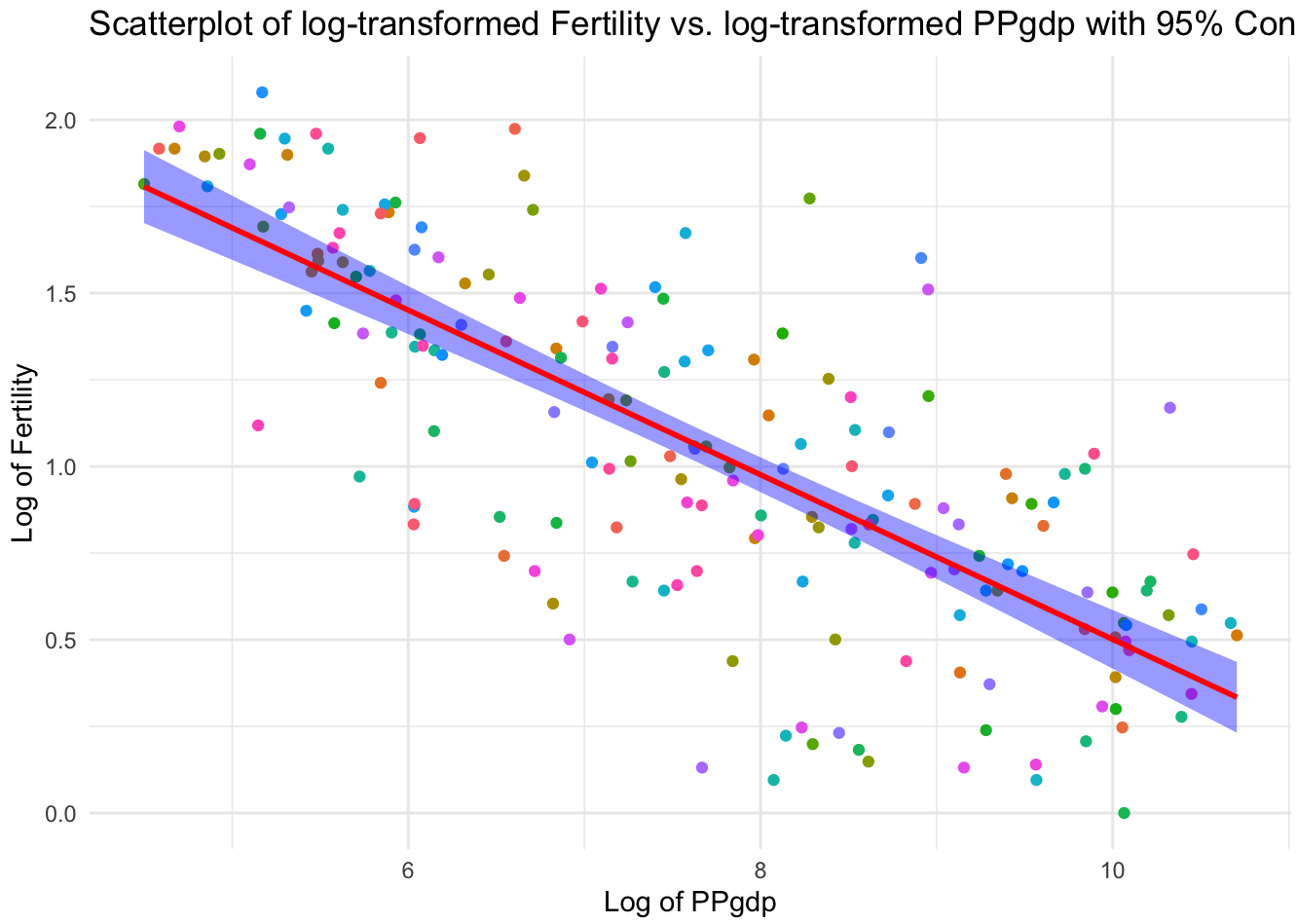
```
library(tidyr)

new_data <- data.frame(log_PPgdp = seq(min(data$log_PPgdp), max(data$log_PPgdp), length.
out = 100)) %>%
  # Add predictions and confidence intervals
  mutate(predictions = predict(linear_model, newdata = .),
         conf_low = predict(linear_model, newdata = ., interval = "confidence", level =
0.95)[,2],
         conf_high = predict(linear_model, newdata = ., interval = "confidence", level =
0.95)[,3])

head(new_data)
```

	log_PPgdp <dbl>	predictions <dbl>	conf_low <dbl>	conf_high <dbl>
1	4.499810	1.807434	1.702060	1.912807
2	4.562489	1.792548	1.688806	1.896291
3	4.625167	1.777663	1.675544	1.879782
4	4.687846	1.762778	1.662274	1.863281
5	4.750525	1.747892	1.648997	1.846788
6	4.813204	1.733007	1.635711	1.830303

6 rows

```
ggplot(data, aes(x = log_PPgdp, y = log_Fertility)) +
  geom_point(aes(color = Locality)) +
  geom_smooth(method = "lm", se = TRUE, color = "red", fill = "blue", level = 0.95, aes
(group = 1)) +
  labs(title = "Scatterplot of log-transformed Fertility vs. log-transformed PPgdp with
95% Confidence Band",
       x = "Log of PPgdp",
       y = "Log of Fertility") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Assuming that the same relationship between Fertility and PPgdp holds, give a 99% prediction interval on Fertility for a region with PPgdp 25,000 US dollars in 2018

```
log_ppgdp_2018 <- log(25000)
# 99% prediction interval
Pred_int <- predict(linear_model, newdata = data.frame(log_PPgdp = log_ppgdp_2018),
                                interval = "prediction", level = 0.99)
print("Prediction Interval")
```

```
## [1] "Prediction Interval"
```

```
print(Pred_int)
```

```
##         fit       lwr     upr
## 1 0.4711468 -0.415426 1.35772
```

```
exp_prediction_interval <- exp(Pred_int)
print("Exp_prediction_interval")
```

```
## [1] "Exp_prediction_interval"
```

```
print(exp_prediction_interval)
```

```
##       fit       lwr      upr
## 1 1.60183 0.6600591 3.887319
```

WE are 99% confident that the actual Fertility rate for a region with a PPgdp of $25,000 in 2018 will fall between approximately 0.660 and 3.887 births per 1000 females.

] Based on the diagnostic plots my concern on the above hypothesis testing and inferences:

The slight deviations in the QQ plot's tail can cause some concerns about the normality assumption despite it not being drastic. There are not a lot of apparent very influential points based on Cook's distance, but we still need to look at the points that are close to the threshold just to make sure they are not disproportionately influencing the model. So overall the diagnostic plots suggest that many of the key assumptions of linear regression are adequately met with minor concerns.

Reference: https://bookdown.org/logan_kelly/r_practice/p09.html (https://bookdown.org/logan_kelly/r_practice/p09.html)