# Homework2

Chandnee das

2/13/2022

Exercise: 1

(a) The assumptions for the simple linear regression model are:

Linearity

Independence

Constant variance

Normality

Plot of the residual vs fitted values is the one kind of diagnostics tool to check these assumptions.

Using the residual plot, we can check the linear regression model has constant variance or nonconstant variance around the fitted line. We can also check is there any outlier or not. After seeing the pattern from the residual vs fitted plot, We can identify the linearity or nonlinearity from the regression model and we can identify outliers and their position from the fitted regression line.

To verify the normality assumption, normal probability plot or QQ plot is one kind of diagnostic tool. In QQ plot, sample residual quantiles are placed on the Y axis and theoretical quantiles are placed on the X axis. If the data points follow a straight line, then we can say that sample data are normally distributed. On the other hand, if the data points deviate from the straight line, that means the sample data are not normally distributed.

(b)

An outlier is a point that deviates from the set of bulk data. An outlier has y value does not follow the pattern compared to the other data in the model.

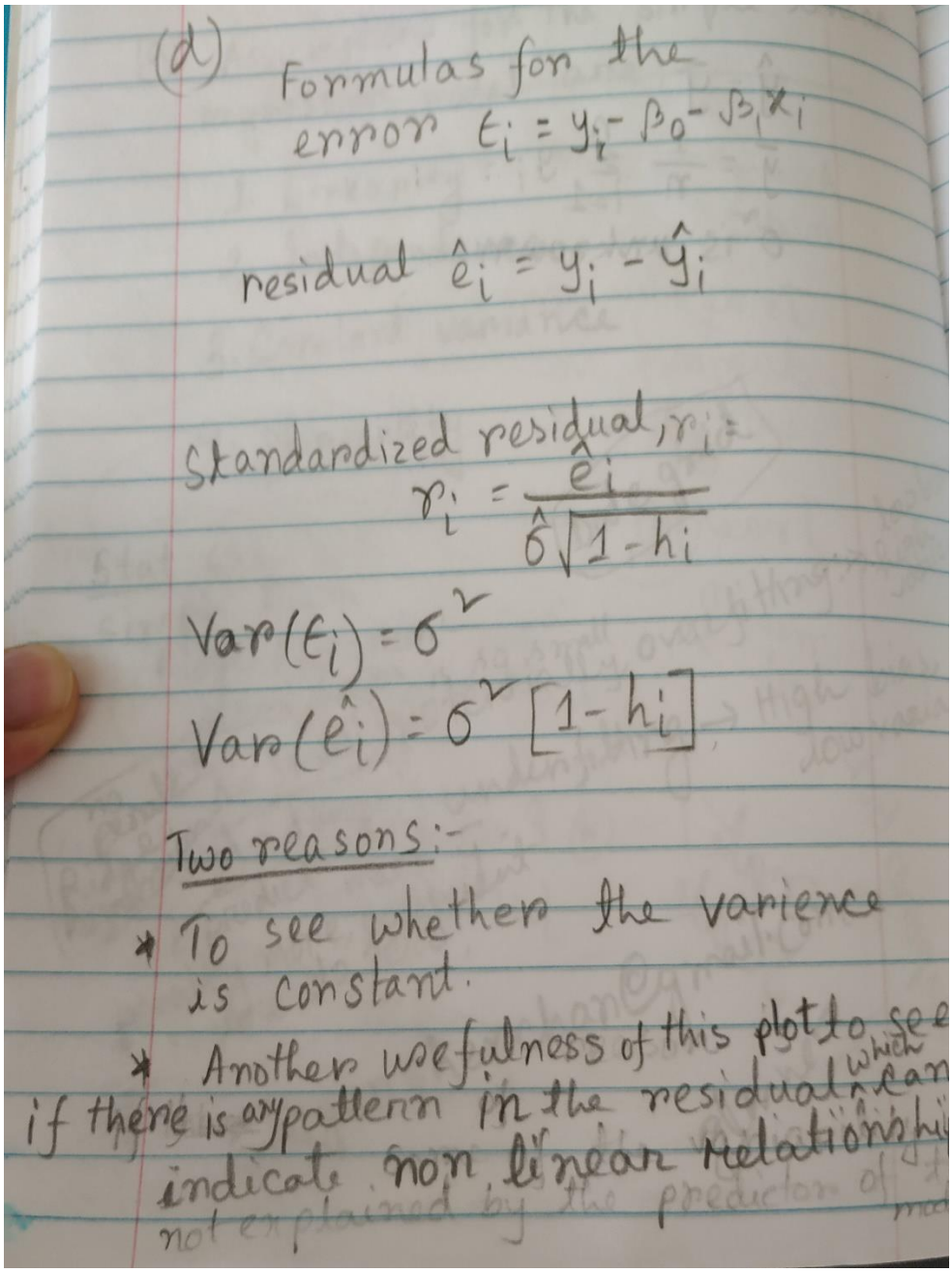
For simple linear regression, the rule is commonly used to classify points as outliers when the data point of standard residuals falls outside the interval from -2 to 2.

(c)

A point which has x values far from the other x values of the model, that means the point is called high leverage point. High leverage point that is also an outlier can change the estimation of least square line.

The rule for the high leverage point of simple linear regression model is hi> 4/n.

(d)



Exercise 2:

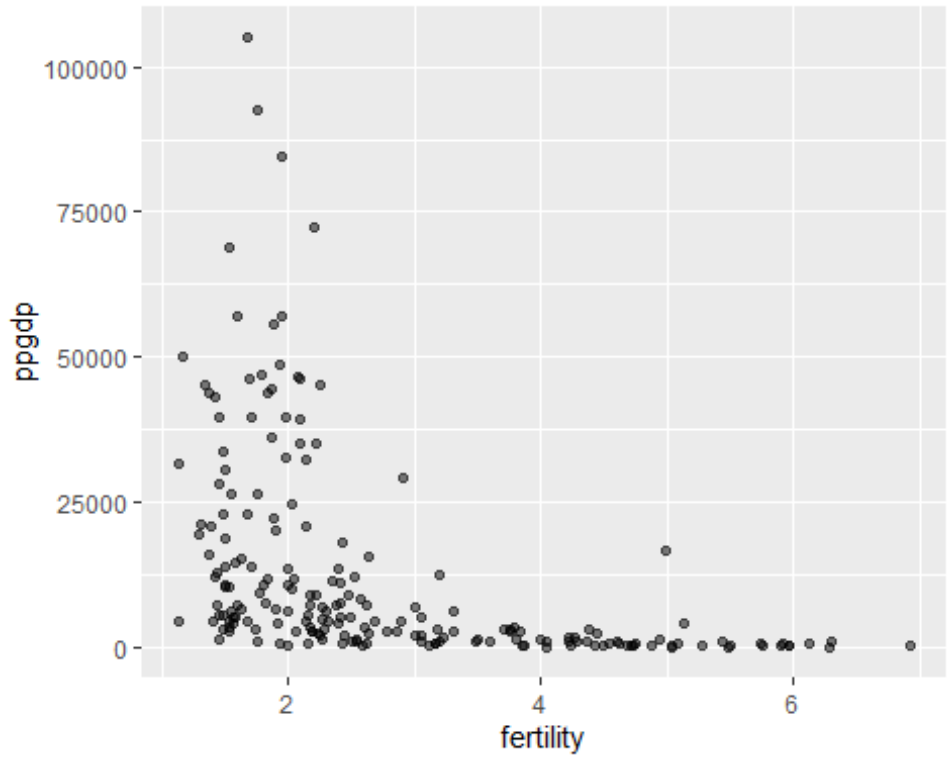
(a) True
(b) False
Answer : The square root transformation used to stabilize the variance for count data.
(c) False
Answer: The transformation can be applied to predictor only or response variable only or both .
(d) True
(e) True

## Exercise 3

```
UN11 <- read.csv("UN11.csv")
```

### (a)

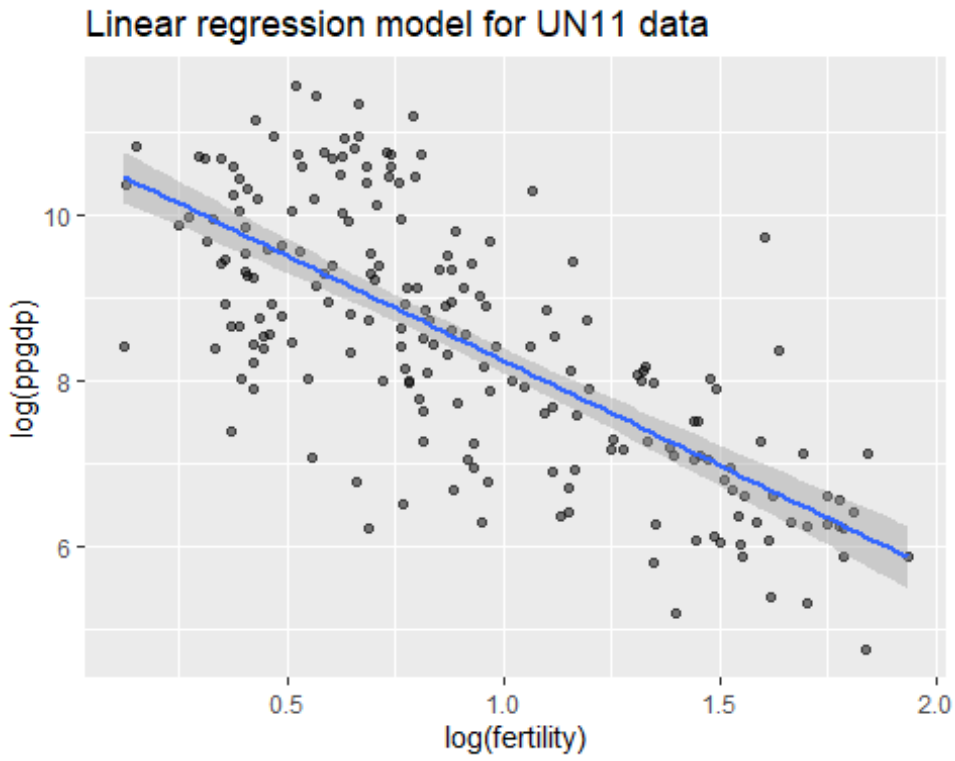
```
library(ggplot2)
ggplot(UN11, aes(fertility, ppgdp)) + geom_point(size=1.5, alpha=0.5)
```




Looks like they dont have linear relationship,thats why we consider log transformation.

## (b)

```
ggplot(UN11, aes(log(fertility), log(ppgdp))) + geom_point(size=1.5,
alpha=0.5) + geom_smooth(method='lm')+ggtitle("Linear regression model for
UN11 data")

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




The assumptions appear reasonably linear.

## (c)

```
lm1<-lm(log(fertility) ~  log(ppgdp), data=UN11)
summary(lm1)

##
## Call:
## lm(formula = log(fertility) ~ log(ppgdp), data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79828 -0.21639  0.02669  0.23424  0.95596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.66551    0.12057   22.11   <2e-16 ***
## log(ppgdp)  -0.20715    0.01401  -14.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3071 on 197 degrees of freedom
## Multiple R-squared:  0.526,  Adjusted R-squared:  0.5236
## F-statistic: 218.6 on 1 and 197 DF,  p-value: < 2.2e-16
```

## (d)

log(fertility)=2.66551-0.20715log(ppgdp)

## (e)

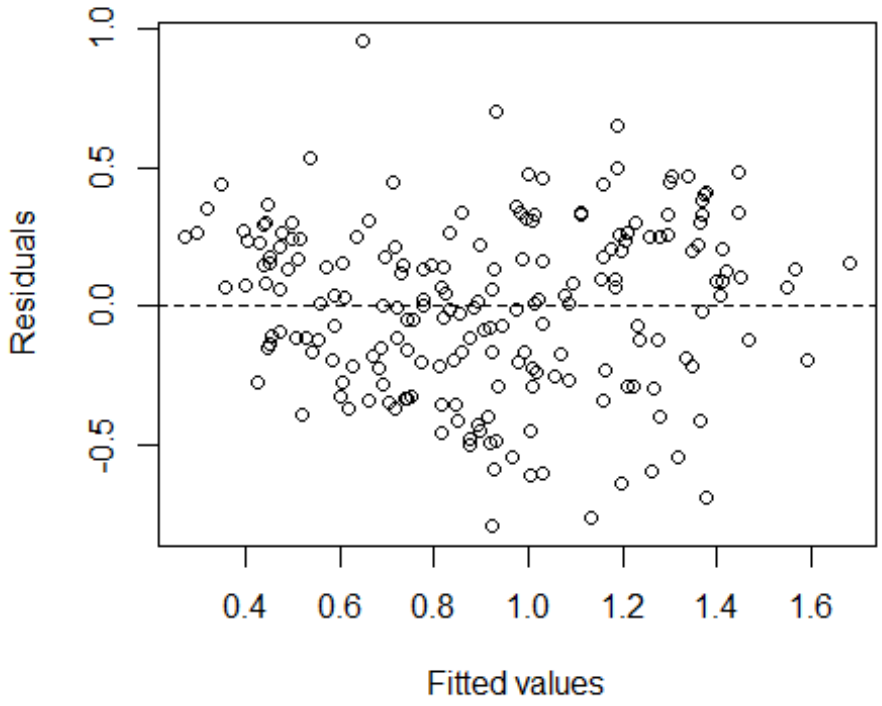
log(fertility) will decrease by slope(-.20715) with the unit increase in log(ppgdp). So the prediction and log of response variable have inverse linear relationship.

## (f)

```
new_x <- data.frame(ppgdp = 1000)
pred<- predict(lm1, newdata = new_x, interval="prediction", level = 0.95)
pred

##        fit       lwr      upr
## 1 1.234567 0.6258791 1.843256

exp(pred)

##        fit      lwr     upr
## 1 3.436891 1.869889 6.31707
```

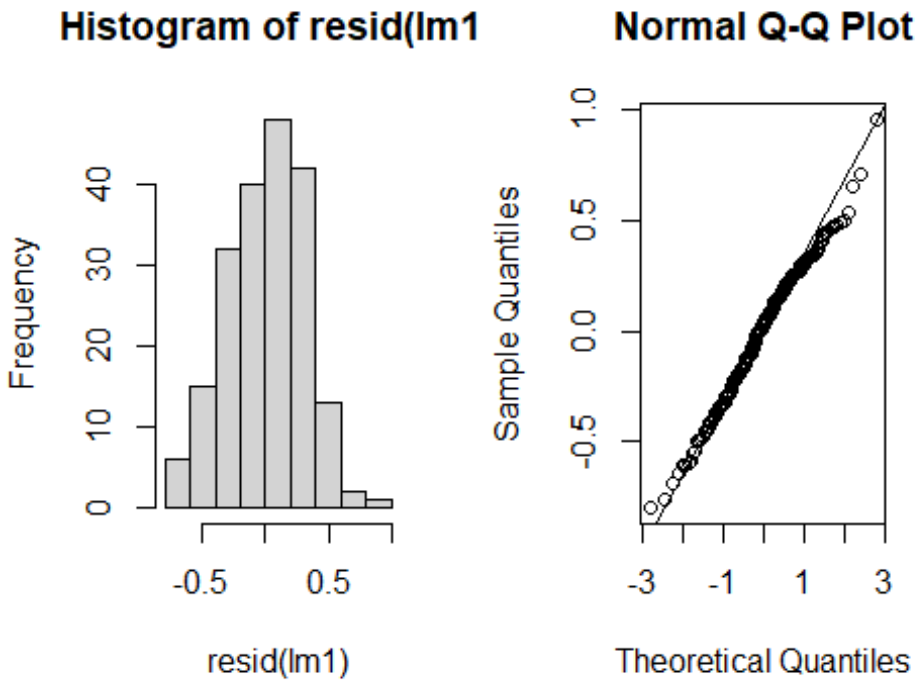
## (g)

```
par(mar=c(4.5, 4.5, 2, 2))
plot(predict(lm1), resid(lm1), xlab='Fitted values', ylab='Residuals')
abline(h=0, lty=2)
```




the variance close to constant and no obvious outliers. So the linearity and constant varience assumptions of linear regression is satisfied.

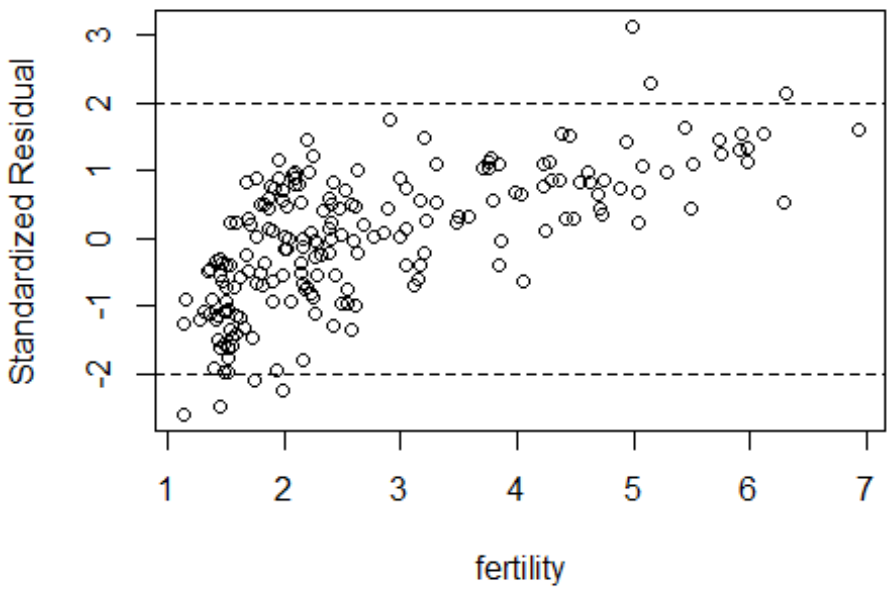
```
par(mfrow=c(1,2))
hist(resid(lm1))
qqnorm(resid(lm1))
qqline(resid(lm1))
```




*the residual close to normal. So the normality assumptions of linear regression is satisfied.*

## (h)

```
plot(UN11$fertility, rstandard(lm1),
xlab = "fertility", ylab = "Standardized Residual")
abline(h=c(-2,2), lty=2)
```




```
ind <- which(abs(rstandard(lm1)) > 2)
UN11_2<-UN11[ind,]
UN11_2$ï..country
```

```
## [1] "Angola"                "Bosnia and Herzegovina" "Equatorial Guinea"
## [4] "Moldova"               "North Korea"            "Viet Nam"
## [7] "Zambia"
```

*I think we do not need to remove the outliers because the stansardized residuals shows a pattern which indicates that some variabilities of the data are not explained by the model and the outliers may not be true outliers.*

## Bonus

$$\hat{y} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$E(\hat{y}^*) = E(\hat{y} \mid x = x^*) = u$$

$$\text{var}(\hat{y}^*) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

$$\text{var}(\hat{y}) = \text{var}(\bar{y}) = \frac{\sigma^2}{n}$$

$$T = \frac{\hat{y}^* - u}{s^2/n}$$

confidence interval $\hat{y}^* \pm t(\alpha/2, n)\, s^2/n$

$$= \bar{y} \pm t(\alpha/2, n)\, s^2/n$$

Prediction interval

$$T = \frac{y^* - \hat{y}^*}{s^2(1 + 1/n)}$$

$$\hat{y}^* \pm t(\alpha/2, n)\, s^2(1 + 1/n)$$

$$= \bar{y} \pm t(\alpha/2, n)\, s^2(1 + 1/n)$$

Prediction interval is wider than confidence interval.