Homework 2 STAT632

Winnie Lu

2/17/2022

Exercise 1

(a)

The assumptions for the simple linear regression model can be explained with the LINE mnemonic: The relationship between the predictor and response variable is linear, the samples are drawn randomly without replacement from independent populations, the data follows a normal distribution and each of the subpopulations (each distinct x-value) has a common or equal variance. One of the most common way to check the normality assumption is by using a QQ plot. A QQ plot is a graphical tool we use to check normality. It plots sample quantiles against theoretical quantiles. If the points follow a straight line, then the data is approximately normal; however, if the points deviate from the straight line, then that indicates a deviation from the normal distribution. Another way to check for LINE assumptions is by using residual plots. If we plot the residuals vs fitted values, we should ideally see a data with points showing no discernible pattern and scattered around 0. This would tell us that the data has equal variances.

(b)

Outliers are points whose y-values don’t follow the other pattern bulk of data points. For a point to be considered an outlier, it would have to be outside of the interval of -2 to 2 in a standardized residual plot.

(c)

Leverage points are points whose x-values don’t follow the other pattern bulk of data points. For a point in question to be a high leverage point, the point, $h_i > 4/n$. This rule only applies to simple linear regression.

(d)

$\hat{e_i} = y_i - \hat{y}_i . \epsilon_i$ and $\hat{e_i}$ are both random errors. ϵ_i are random errors from the population, while $\hat{e_i}$ are random errors from the sample. If sampling is done well, then $\hat{e_i}$ should behave in the same way as ϵ_i since we are unable to measure ϵ_i.

$\mathrm{Var}\left(\hat{e}_i\right) = \sigma^2 \left[1 - h_i\right]$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$.

Even though the residuals don’t have constant variance, our errors do. This is one of the reasons why it’s useful to examine a standardized residual plot; it’ll allow us to view high leverage points in a data set. If there are not high leverage points, then a regular residual plot will not look much different from a standardized residual plot. Another reason why it’s useful to have the standardized residual (vs fitted) plot is that it will allow us to view outliers more clearly and we can tell how far the outliers are from the fitted regression line.

Exercise 2

(a)

True, residual plots are useful in assessing linearity and constant variance

(b)

False, the square root transformation is more commonly applied to count data to stabilize variance. Log transformations are generally applied to skewed data that ranges over several magnitudes.

(c)

False, transformations can be applied to the response variable, the predictor or both; it doesn’t have to always be applied to both variables.

(d)

False, as we have seen with Anscombe’s plots, even if the $R^2 = 1$, it does not mean that all the data points fit a straight line. We have to examine the data points itself to conclude if the line is linear and if R^2 is relevant. Some R^2 could be close to 1, but have a non-linear relationship.

(e)

True; transformations are used to linearize the relationship between X and Y in SLR and overcome problems from having a non-constant variance.

Exercise 3

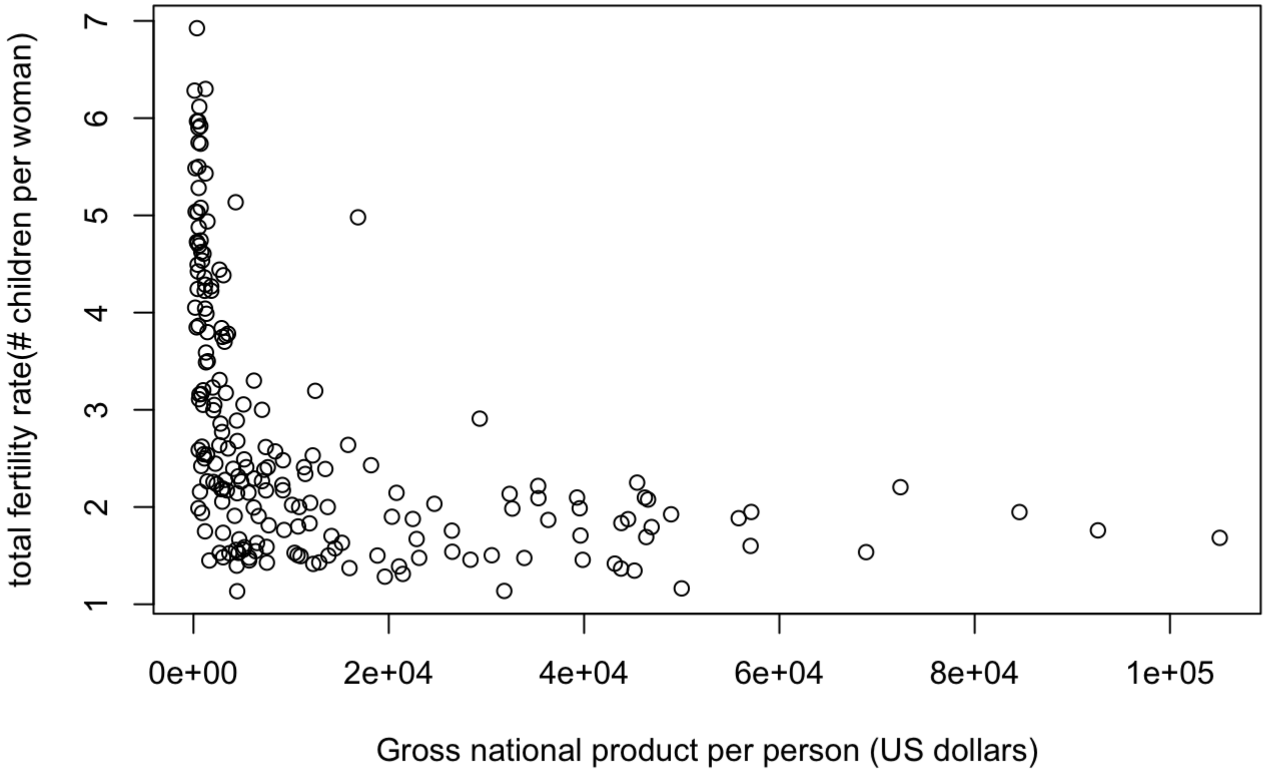
```
UN11 <- read.csv("~/Library/Mobile Documents/com~apple~CloudDocs/stat632/Hmwk/UN11.csv")
head(UN11)
```

```
##       country     region fertility   ppgdp lifeExpF pctUrban
## 1 Afghanistan       Asia     5.968   499.0    49.49       23
## 2     Albania     Europe     1.525  3677.2    80.40       53
## 3     Algeria     Africa     2.142  4473.0    75.00       67
## 4      Angola     Africa     5.135  4321.9    53.17       59
## 5    Anguilla  Caribbean     2.000 13750.1    81.10      100
## 6   Argentina Latin Amer     2.172  9162.1    79.89       93
```

(a)

As shown from the graph, we would need to consider a log transformation for this data since the original scatterplot shows that the data is extremely skewed (right skewed).

```
plot(fertility ~ ppgdp, data=UN11, xlab="Gross national product per person (US dollars)"
, ylab="total fertility rate(# children per woman)")
```

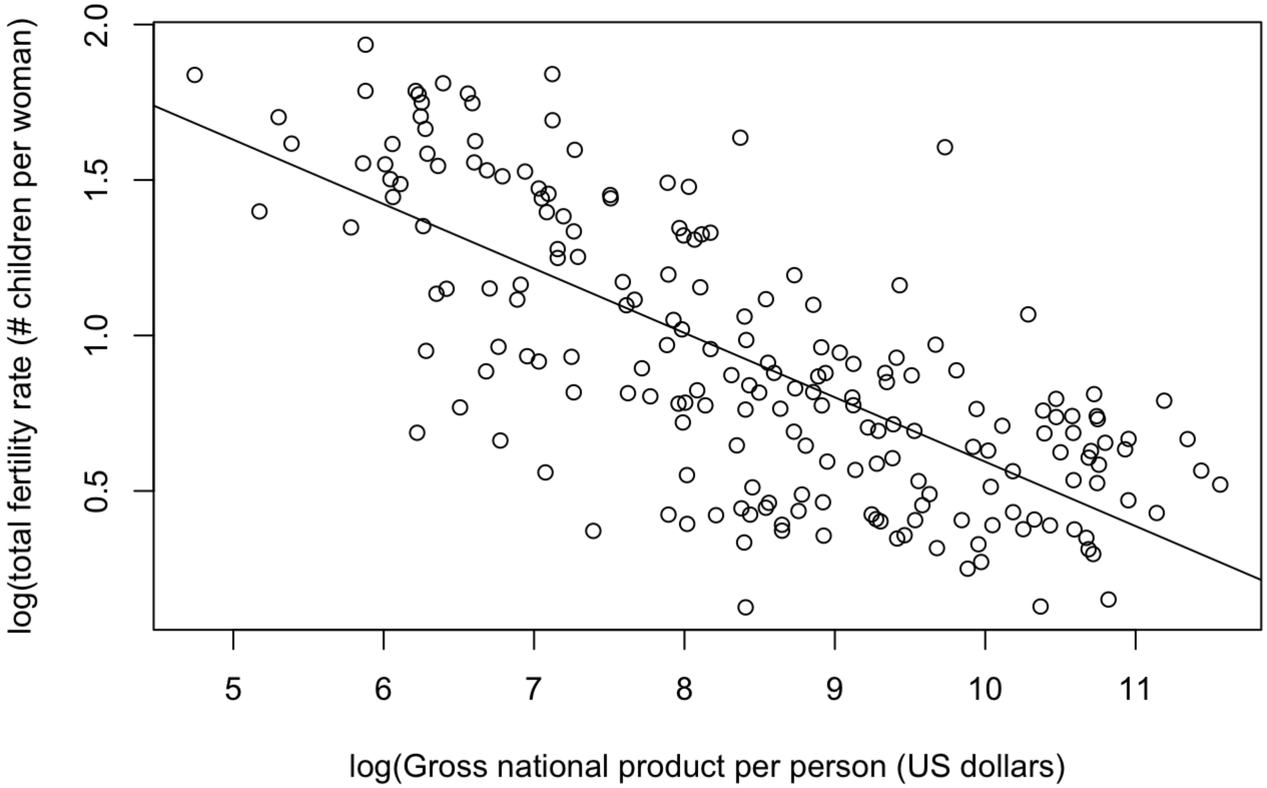


(b)

With the log transformation, the previously skewed data now appears to be reasonably linear. We see a negative linear relationship between x and y.

```
plot(log(fertility) ~ log(ppgdp), data=UN11, xlab="log(Gross national product per person
(US dollars)", ylab="log(total fertility rate (# children per woman)")

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



(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)

$$log(\widehat{fertility}) = 2.665 - 0.2071 log(ppgdp)$$

(e)

A unit increase in log(ppgdp) is associated with a decrease in log(fertility) by -0.2071.

(f)

For a gross national product per person of $1000, we would expect to see a 1.234 log(fertility) rate. We would be 95% confident that the log(fertility) rate would lie between 0.625 and 1.84.

After exponentiation, for a gross national product per person of $1000, we would expect to see a 3.436 fertility rate. We would be 95% confident that the fertility rate would lie between 1.869 and 6.317.

```
2.665-.2071*log(1000)
```

```
## [1] 1.234404
```

```
#this calculation is done with natural log!
#When we exponentiate both sides, we get Fertility = 3.436

#prediction for log(fertility) when ppgdp=1000:
pred <-predict(lm1, interval="prediction", data.frame(ppgdp=1000))
pred
```

```
##        fit       lwr      upr
## 1 1.234567 0.6258791 1.843256
```

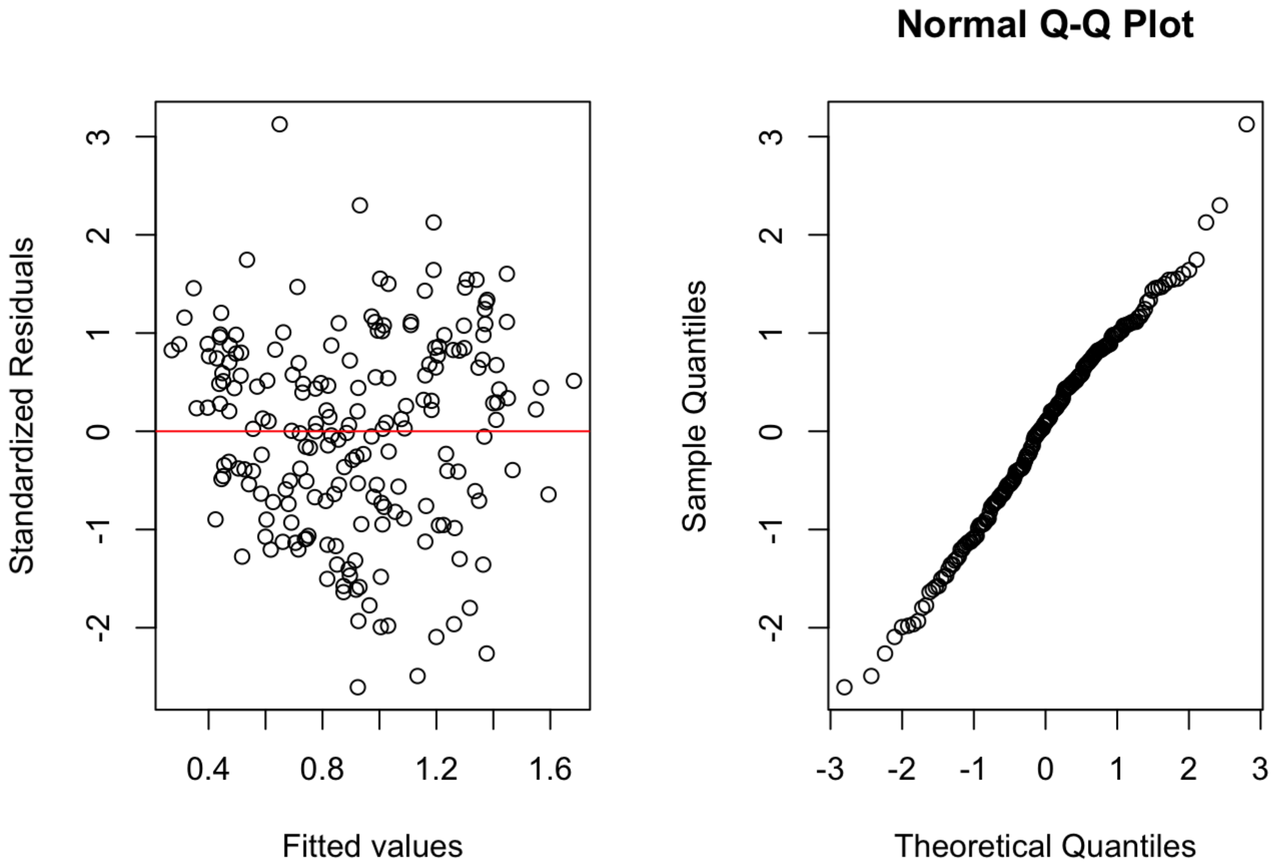
```
#exponentiate to make prediction for fertility:
exp(pred)
```

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

(g)

After taking a log transformation of our data, we can see that the assumptions of linearity and constant variance have been reasonably satisfied. In the residual vs fitted plot, the data is scattered with no discernible pattern and are scattered about 0. In the QQ plot, we see that the data points reasonably follow a straight line indicating normality.

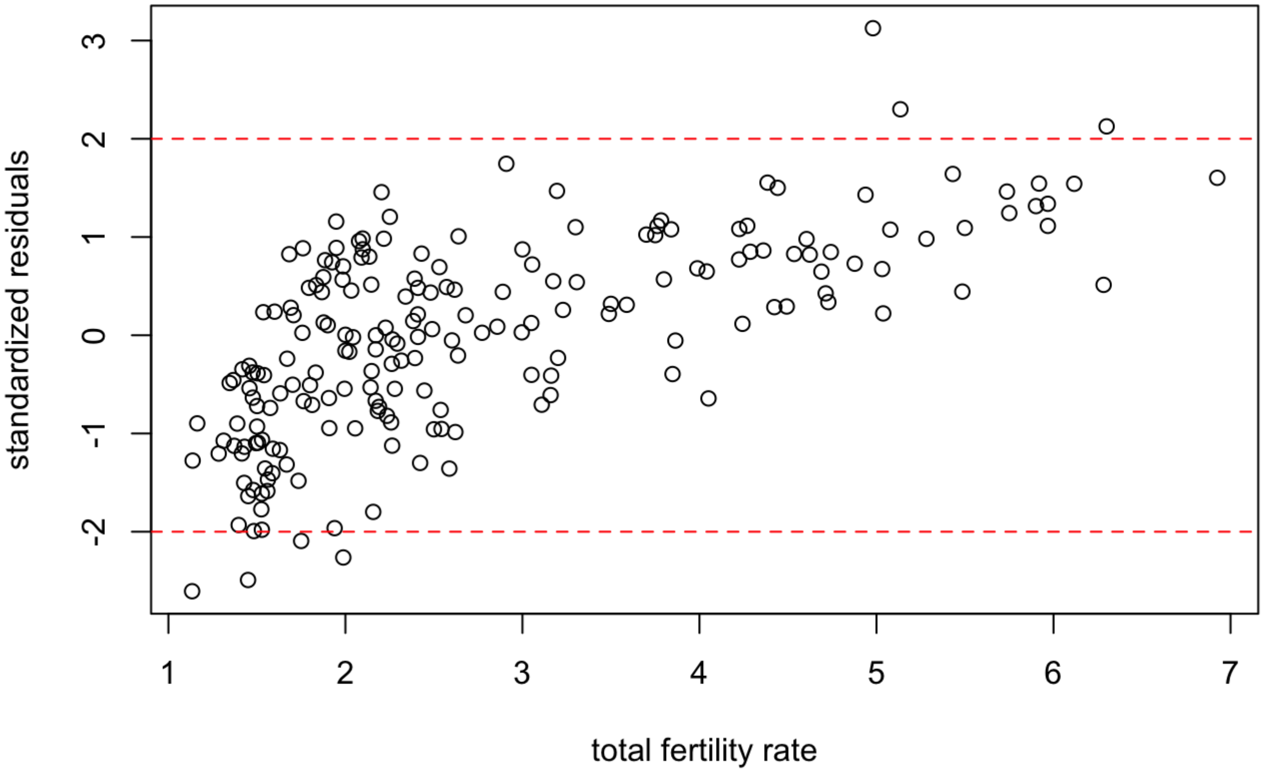
```
par(mfrow=c(1,2))
plot(predict(lm1), rstandard(lm1), xlab="Fitted values", ylab="Standardized Residuals")
abline(h=0, col="red")
qqnorm(rstandard(lm1))
```



(h)

Since outliers are data points whose y-values deviate, we have to examine the response variable. The countries that are considered outliers are Angola, Bosnia & Herzegovina, Equatorial Guinea, Moldova, North Korea, Vietnam, and Zambia. No, I do not believe it to be necessary to remove the outliers because it does contribute information to our model. There could be other factors affecting fertility rate, including, but not limited to: socioeconomic status, political agenda (laws that limit childbirth rate), intrinsic health issues, etc.

```
plot(UN11$fertility, rstandard(lm1), xlab="total fertility rate", ylab="standardized residuals")
abline(h=c(-2,2), lty=2, col="red")
```



```
#identify outliers
ind <-which(abs(rstandard(lm1))>2)
UN11[ind,]
```

```
##                      country region fertility   ppgdp lifeExpF pctUrban
## 4                     Angola Africa     5.135  4321.9    53.17       59
## 23  Bosnia and Herzegovina Europe     1.134  4477.7    78.40       49
## 58       Equatorial Guinea Africa     4.980 16852.4    52.91       40
## 118                Moldova Europe     1.450  1625.8    73.48       48
## 134            North Korea   Asia     1.988   504.0    72.12       60
## 196               Viet Nam   Asia     1.750  1182.7    77.44       31
## 198                 Zambia Africa     6.300  1237.8    50.04       36
```