Stat632HW2

Brandon Keck

Exercise 1.

(a)

The assumptions for the simple linear regression model are:

1. Linearity
2. Independence
3. Constant Variance
4. Normality

Two diagnostics that are commonly used to check assumptions:

1. One useful diagnostic is a plot of the residuals versus the fitted values.
2. Another useful diagnostic is to determine whether the proposed regression model is a valid model i.e. determine whether it provides an adequate fit to the data.

(b)

For a point to be considered an outlier it is a point that does not follow the majority of the data. What this means is that an outlier will have y-values that do not follow the pattern of the data.

****The rule**** for identifying outliers is a point whose standardized residual falls outside the interval from -2 to 2.

(c)

For a point to have **high leverage** it is considered an extreme value of a variable within a dataset.

The rule that identifies a point as a high leverage point is a point that has $h_i > 4/n$

(d)

Error formula: $\epsilon i = Y_i - E(Y_i)$

Residuals formula: $\hat{e_i} = y_i - \hat{y_i}$

Standardized Residual : $r_i = \frac{\hat{e_i}}{\hat{\sigma}\sqrt{1-h_i}}$

Variance of Errors: $Var(\epsilon_i) = \sigma^2$

Variance of Residuals: $Var(\hat{e_i}) = \sigma^2[1-h_i]$

It is useful to look at the standardized residuals over the fitted values when there are points of high leverage in the data set.

However, if there are no points of high leverage, then generally speaking there is little difference between the plot of raw residuals and the standardized residuals.

Exercise 2.

(a)

True

(b)

False - Log transformation is commonly applied to skewed data that ranges over several ordersof magnitude.

(c)

False - Transformations may be applied to either the response variable (Y), or the explanatory variable (X) or both. It is not necessary to have to always transform both variables.

(d)

False - A high R^2 only means the model explains a large proportion of variance in Y, but it does not guarantee a good fit.

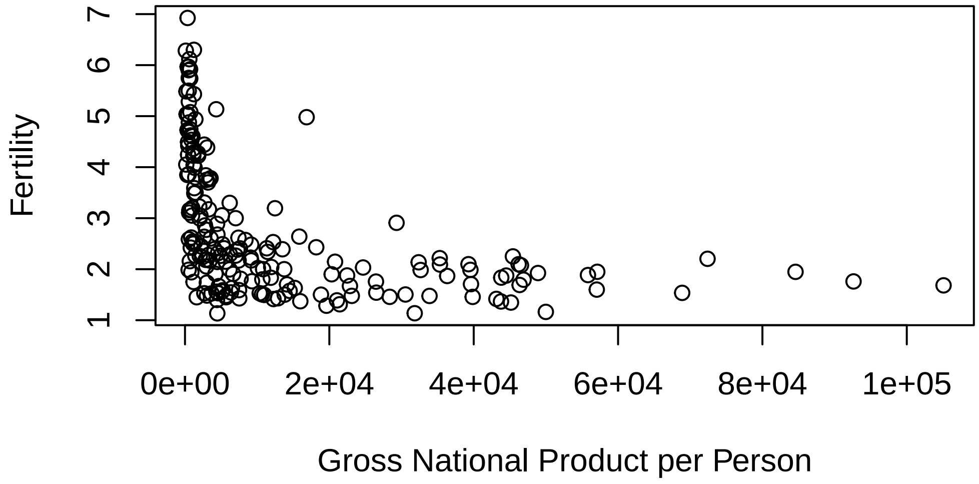
(e)

True

Exercise 3.

```
library(readr)
UN11 <- read_csv("~/Documents/EastBay/Spring2025/Stat632Regression/UN11.csv", show_col_typ

# head(UN11)
# create the linear model with lm
lm1 <- lm(fertility ~ ppgdp, data = UN11)
plot(fertility ~ ppgdp, xlab = "Gross National Product per Person",
     ylab = "Fertility", data = UN11)
```



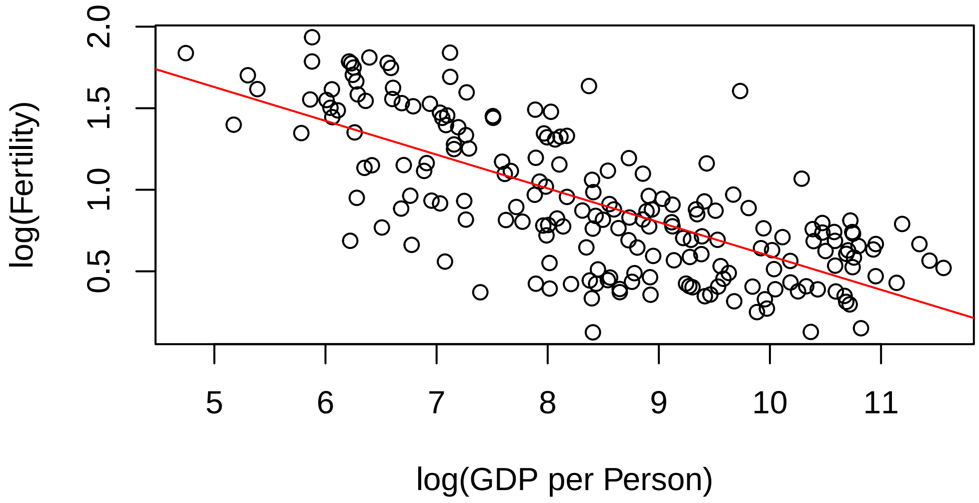
```
# plot scatter plot with plot function
```

The reason why we should consider using a log transformation is because from the scatter plot we observe that the predictor variable is skewed right. That being said the log transformation is commonly applied to skewed data the ranges over several order of magnitude.

(b)

```
# create the linear model with lm
lm2 <- lm(log(fertility) ~ log(ppgdp), data = UN11)
plot(log(fertility) ~ log(ppgdp), xlab = "log(GDP per Person)",
     ylab = "log(Fertility)", data = UN11)
# plot scatter plot with plot function

abline(lm2, col = "red")
```



After taking the log of both the explanatory (X) and response variables (Y) we now observe a reasonable negative association of linearity between the two variables.

(c)

```
# create the linear model with lm
lm2 <- lm(log(fertility) ~ log(ppgdp), data = UN11)

summary(lm2)
```

```
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
```

```
# print out the summary statistics
```

(d)

Equation of linear model:

$log(\widehat{fertility}) = \hat{\beta}_0 + \hat{\beta}_1 * (log(ppgdp))$

= 2.66551 - 0.20715 x (log(ppgdp))

(e)

Make a prediction for log(fertility) when ppgdp is $46545.9:

```
2.66551 - 0.20715 * log(46545.9)
```

```
[1] 0.4390216
```

0.439 (log $)

let's exponentiate both sides to get the prediction for fertility rate:

```
exp(0.4390216)
```

```
[1] 1.551189
```

Make a prediction for log(fertility) when ppgdp is $46545.9:

```
pred <- predict(lm2, data.frame(ppgdp = 46545.9),
                interval = "prediction")
pred
```

```
        fit        lwr      upr
1 0.4390212 -0.1714187 1.049461
```

```
exp(pred)
```

```
       fit       lwr      upr
1 1.551188 0.8424688 2.856111
```

Interpretation: For every 1% increase in GDP per capita, the fertility rate decreases by 20.7%.

(f)

```
df1 <- predict(lm2, data.frame(ppgdp = 1000),
               interval = "predict")
df1
```

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

If the log(ppgdp) were 1,000, we would predict that the log(fertility) would be approximately 1.235. With a 95% probability log(fertility) will fall between 0.626 and 1.843 for a log(ppgdp) of 1000.

```
exp(df1)
```

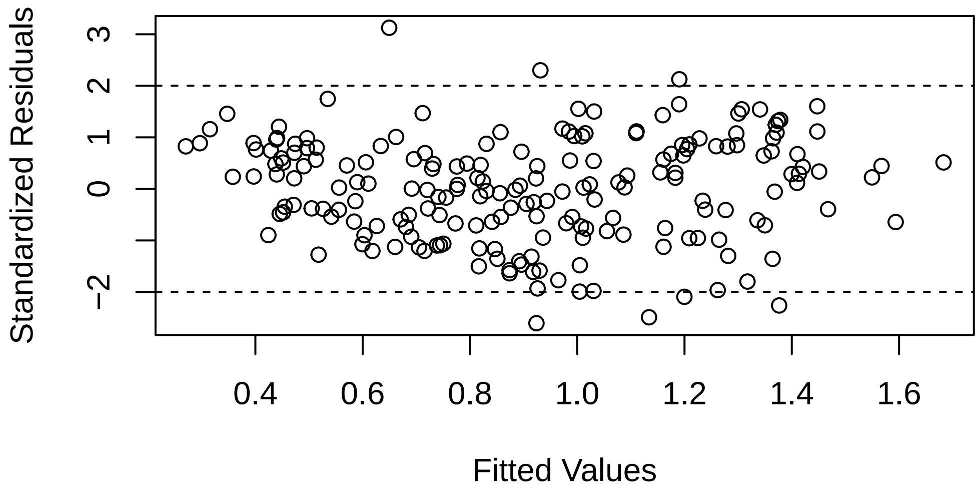
```
       fit      lwr     upr
1 3.436891 1.869889 6.31707
```

For a country with GDP per capita = 1000:

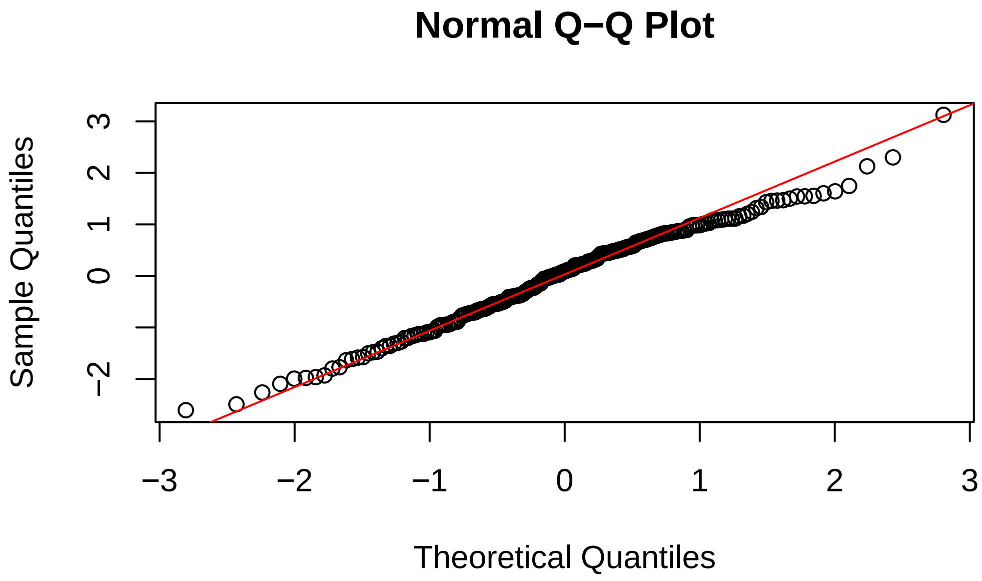
The predicted fertility rate is approxmiately 3.44 children per women. With a 95% prediction interval this suggests that the rate could range from 1.87 to 6.32 children per women. Given the variability in the data what this tells us is that a country with GDP of 1000 could have fertility rates anywhere within this range.

(g)

```
# Standardized residuals vs fitted values
plot(predict(lm2), rstandard(lm2),
     xlab = "Fitted Values", ylab = "Standardized Residuals")
n <- nrow(UN11)
abline(h=c(-2,2), lty = 2) # Threshold for outliers
```



```
# Q-Q plot of Normality
qqnorm(rstandard(lm2)) # QQ Plot
qqline(rstandard(lm2), col = "red")
```



From the standardized residuals versus fitted values plot, and the Normal Q-Q plot there are some noticeable outliers. Because of these outliers we should investigate further in order to be assured of our model. There is also some deviation near both ends of the tails of the Q-Q plot. However, all conditions/assumptions seem to be satisfied.

(h)

```
#identify outliers
ind <- which(abs(rstandard(lm2)) > 2)
UN11[ind, ]
```

```
# A tibble: 7 x 6
  country                region fertility  ppgdp lifeExpF pctUrban
  <chr>                  <chr>      <dbl>  <dbl>    <dbl>    <dbl>
1 Angola                 Africa      5.14  4322.     53.2       59
2 Bosnia and Herzegovina Europe      1.13  4478.     78.4       49
3 Equatorial Guinea      Africa      4.98 16852.     52.9       40
4 Moldova                Europe      1.45  1626.     73.5       48
5 North Korea            Asia        1.99   504      72.1       60
6 Viet Nam               Asia        1.75  1183.     77.4       31
7 Zambia                 Africa      6.3   1238.     50.0       36
```

The countries that are flagged as outliers are; Angola, Bosnia and Herzegovina, Equatorial Guinea, Moldova, North Korea, Vietnam, and Zambia. The reason they are flagged as outliers is because they fall outside the (-2,2) interval. It is never a good idea to just go and remove outliers. We should investigate further and see if removing them adds any significance to our model. This is a fairly large data set and there are only 7 countries that are considered outliers. One thing that we might consider doing is to change our threshold to another value say (-3,3).