HW2_632

Prashan A. Welipitiya

2/14/2020

Exercise 1

a. Linearity, Independence, Normality, and Equal Variance. We check if the residuals vs fitted values has constant variance. We can also check if the relationship between x and y is linear.
b. Outliers are points that dont follow the bulk of the data. In SLR we identify if the interval is outside -2 to 2 or -4 to 4.
c. Leverage points are points with an x-value that are distant from other x values. In SLR we calculate it using $h_i > 4$/n.
d. error - $e_i \sim N(0, \sigma^2)$ residual - $\hat{e}_i = y_i - \hat{y}_i$ standardized residual - $r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_i}}$ $Var(e_i) = \sigma^2$ $Var(\hat{e}_i) = \sigma^2[1 - h_i]$

The residuals vs fitted values clearly shows if there is any unequal variance in the residuals. Also if there are no obvious patterns, then assumptions are reasonably satisfied.

Exercise 2

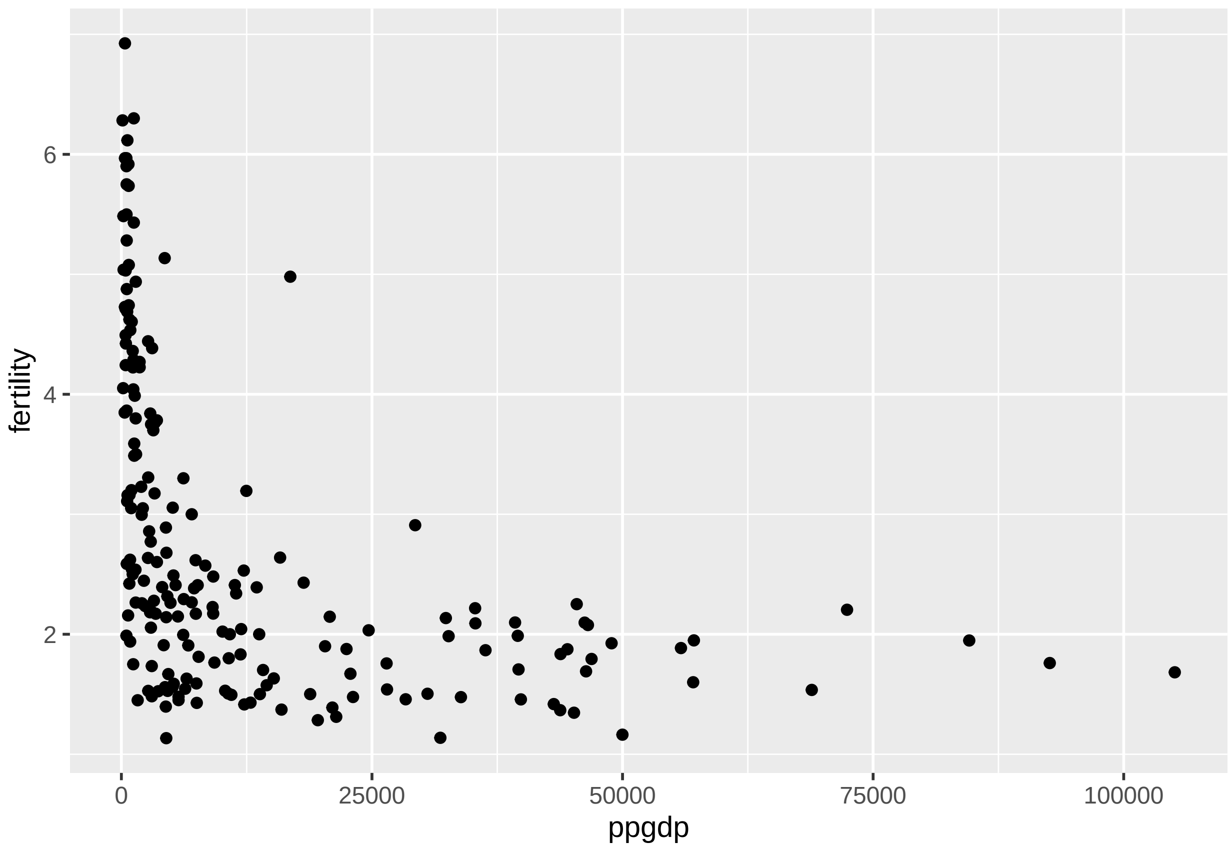
a. False, just constant variance.
b. True
c. False, transforming one variable works just fine.
d. True
e. True

Exercise 3

```r
UN11 <- read.csv("https://ericwfox.github.io/data/UN11.csv")
library(ggplot2)
```

a.

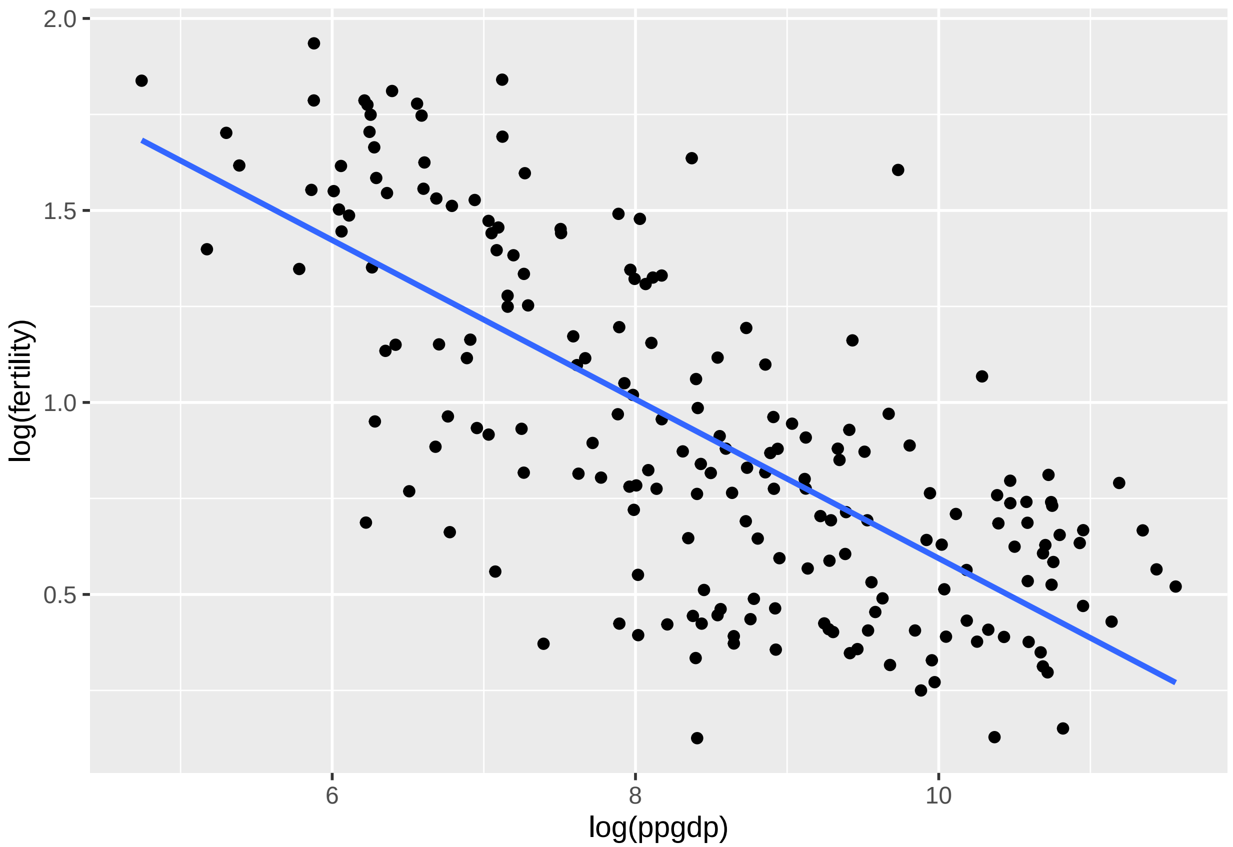
```r
ggplot(UN11, aes(x = ppgdp, y = fertility)) + geom_point()
```



We should consider a log transformation because of how heavily right skewed the plot is.

b.

```
ggplot(UN11, aes(x = log(ppgdp), y = log(fertility))) + geom_point() + geom_smooth(method = "lm", se =
```



In my opinion the association appears to be 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. $log(\hat{y}) = -.2log(\hat{x}) + 2.66$

e. An increase in gross national product per person in US dollars can be associated with a decrease of fertility rate by .2.

f.

```
new_x = data.frame(ppgdp = 1000)
predict(lm1, newdata = new_x, interval="prediction")
```

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

```
exp(1.234567)
```

```
## [1] 3.43689
```

```
exp(0.6258792)
```

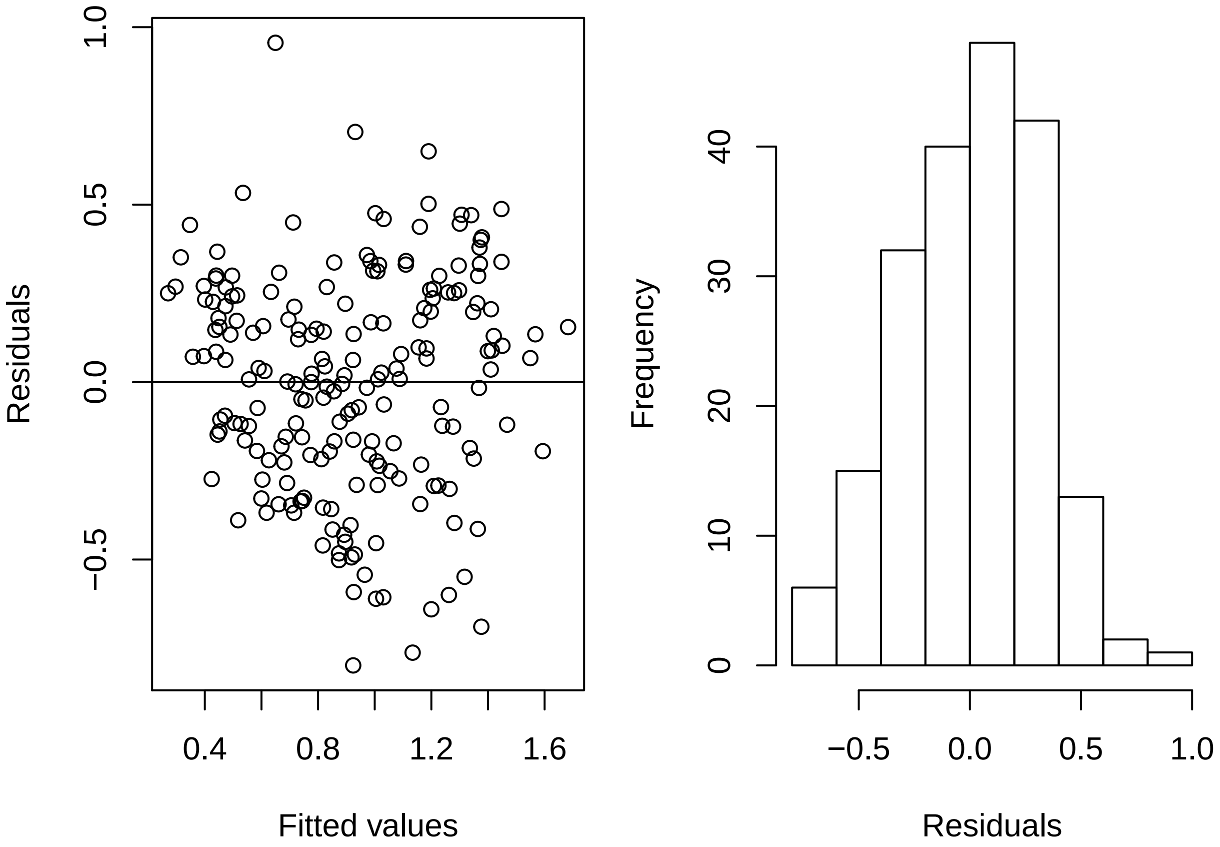
```
## [1] 1.869889
```

```
exp(1.843256)
```

```
## [1] 6.317073
```

g.

```
par(mar=c(4,4,1,1), mfrow=c(1,2))
plot(predict(lm1), resid(lm1), xlab="Fitted values", ylab="Residuals")
abline(h=0)
hist(resid(lm1), main="", xlab="Residuals")
```



The residuals have a normal distribution and the variance looks constant.

h.

```
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
```

I don’t think they need to be removed because it is already linear.