Homework 2

Animesh Sengupta

9/19/2022

```
setwd("/Users/animeshsengupta/Work Directory/DACSS/STAT625/Homeworks")
library(alr4)  # loads the installed package into the workspace so you can use it
```

```
## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
library(summarytools)
library(ggplot2)
```

Answer 1

Answer 2.2.1 : The line y=x essentially means that for cities, the change in price of rice has been constant. If a point lies above the line then that means there has been a rise in price of rice between 2003 and 2009 and if the point is lower then the price of rice has decreased.

Answer 2.2.2 : Vilnius has largest increase in rice price. While mumbai has the largest decrease in rice price.

Answer 2.2.3: if

$$\hat{\beta}_1 < 1$$

generally means that the y value will be lesser than x value. In this case price of 2009 will be lesser than 2003. But the price of rice in 2009 is also determined by other parameter estimate

$$\hat{\beta}_0$$

which can increase the y value from x value. So we cant say for all values of x (i.e price of rice at 2003) is greater than all values of y(price of rice at 2009)

Answer 2.2.4: Fitting linear regression to this might not be appropriate because: 1. A lot of the data points are clustered around one area thus making it harder to accurately draw a model. 2. There are a lot of datapoints with extremeties, with higher values lying as outlier and lower values clustered in a region. This may restrict the model estimation and a log transformation might help.

Answer 2

Answer 2.3.1: A log transformation makes the distribution looks more linearly spread across both the axes. The log transfomation also helps in taking care of extreme values and distributes them linearly across the graph. This linear distribution would make simple linear regression estimations easier.

Answer 2.3.2: b1 essentially captures the rate of growth, hence if it is greater than 1 , then it would lead to exponential growth and if it is equal to one then it would be linear growth and if less than 1 then slower growth. Meanwhile b0 is like an scaling multiplier to the growth function.

Answer 4

Answer 2.15.1 and 2.15.2

```
colnames(wblake)
```

```
## [1] "Age"    "Length" "Scale"
```

```
dim(wblake)
```

```
## [1] 439   3
```

```
summary(wblake)
```

```
##       Age            Length          Scale
##  Min.   :1.000   Min.   : 55.0   Min.   : 1.054
##  1st Qu.:2.500   1st Qu.:138.5   1st Qu.: 3.571
##  Median :5.000   Median :194.0   Median : 5.786
##  Mean   :4.203   Mean   :193.0   Mean   : 5.864
##  3rd Qu.:6.000   3rd Qu.:252.0   3rd Qu.: 8.018
##  Max.   :8.000   Max.   :362.0   Max.   :14.710
```

```
wb<-lm(Length~Age,wblake)
newdat<-data.frame(Age=c(2,4,6))
p2<-predict(wb,newdat,interval="prediction")
p2
```

```
##        fit       lwr      upr
## 1 126.1749  69.73151 182.6184
## 2 186.8227 130.45720 243.1882
## 3 247.4705 191.05332 303.8877
```

```
p3<-predict(wb,data.frame(Age=c(9)),interval="prediction")
p3
```

```
##        fit      lwr      upr
## 1 338.4422 281.7056 395.1788
```

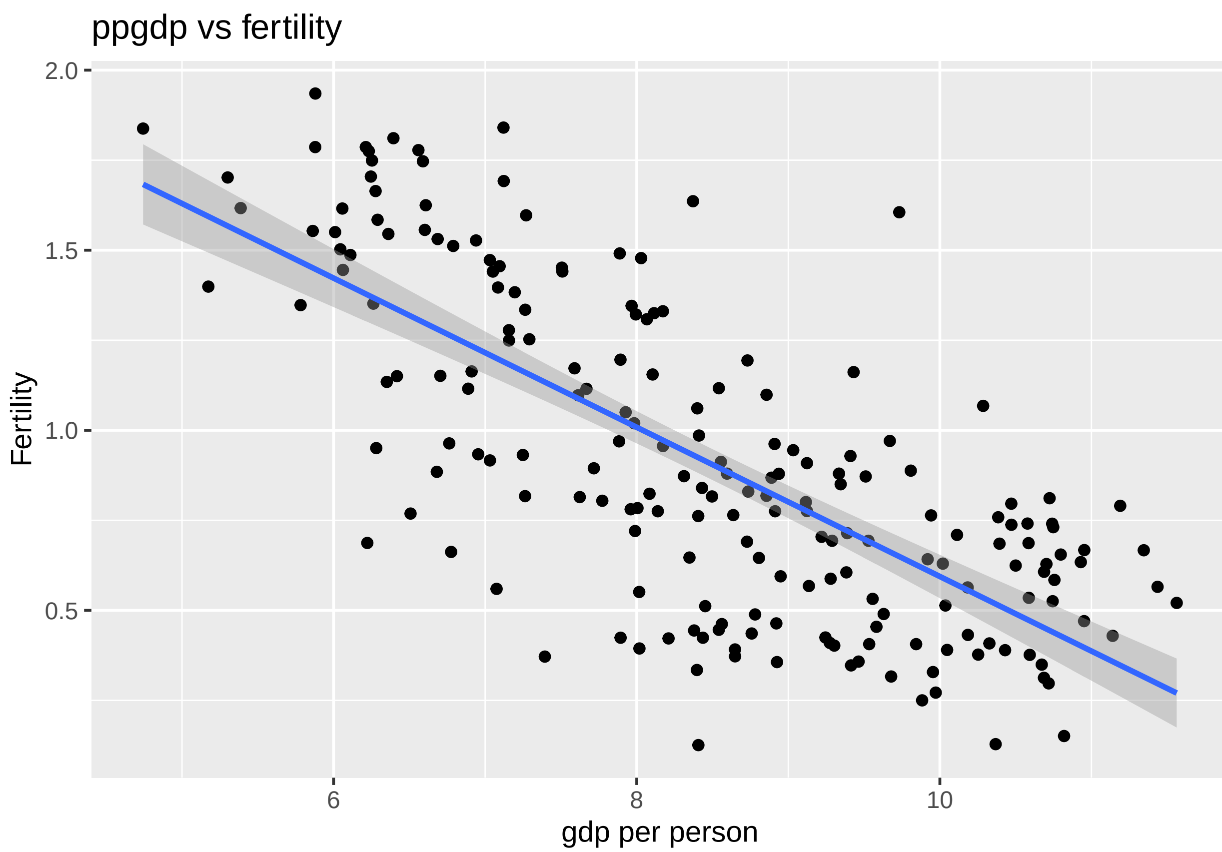
The max age is 8, we are trying to predict for 95% interval for mean age =9 , there are no datapoints for this range hence it can be untrustworthy.

Answer 5

Answer 2.16.1 and 2.16.2

```
unml<-lm(log(fertility)~log(ppgdp),UN11)
b3<- ggplot(UN11,aes(x=log(ppgdp), y=log(fertility))) +
    geom_point()+
    stat_smooth(method="lm")+
    xlab("gdp per person")+
    ylab("Fertility")+
    ggtitle("ppgdp vs fertility")
b3
```

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



answer 2.16.3 and 2.16.4

```
summary(unml)
```

```
##
## 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
```

The t test for slope = 0 , th ep value computed is p<2X10^-16 which is very small and thus very less probable. so we reject the null hypothesis that the slope ==0 . The significance level for this test would be 0.05 as default.

Coefficient of determination is 0.526, means that 52.6% of the variation in fertility can be explained by the ppgdp.

Answer 2.16.5

```r
p1<-predict(unml,data.frame(ppgdp=c(1000)),interval="prediction")
p1
```

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

```r
exp(p1[2])
```

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

```r
exp(p1[3])
```

```
## [1] 6.31707
```

so prediction interval for fertility is (1.87, 6.32)

###Answer 6

The prediction interval of new value y star will be more than the confidence interval of E(y star|x star).

Answer 9

```r
#Answer 2.13.1
summary(Heights)
```

```
##     mheight         dheight
##  Min.   :55.40   Min.   :55.10
##  1st Qu.:60.80   1st Qu.:62.00
##  Median :62.40   Median :63.60
##  Mean   :62.45   Mean   :63.75
##  3rd Qu.:63.90   3rd Qu.:65.60
##  Max.   :70.80   Max.   :73.10
```

```
m9<-lm(dheight ~ mheight, data=Heights)
summary(m9)
```

```
##
## Call:
## lm(formula = dheight ~ mheight, data = Heights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.397 -1.529  0.036  1.492  9.053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.91744    1.62247   18.44   <2e-16 ***
## mheight      0.54175    0.02596   20.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.266 on 1373 degrees of freedom
## Multiple R-squared:  0.2408, Adjusted R-squared:  0.2402
## F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

```
#answer 2.13.2
confint(m9,level=0.99)
```

```
##                  0.5 %     99.5 %
## (Intercept) 25.7324151 34.1024585
## mheight      0.4747836  0.6087104
```

```
#answer2.13.3
p9<-predict(m9,data.frame(mheight=64),level=0.99,interval="prediction")
p9
```

```
##        fit      lwr      upr
## 1 64.58925 58.74045 70.43805
```

2.13.1: The T test for hypothesis b1=0 has a very small p value , thus we can reject this hypothesis and can say that the b1 has some value. 2.13.2 the 99% confidence interval value for b1 is 0.608 2.13.3 Best fir prediction is 64.589

Answer 10

```
#answer 2.4.1
summary(UBSprices)
```

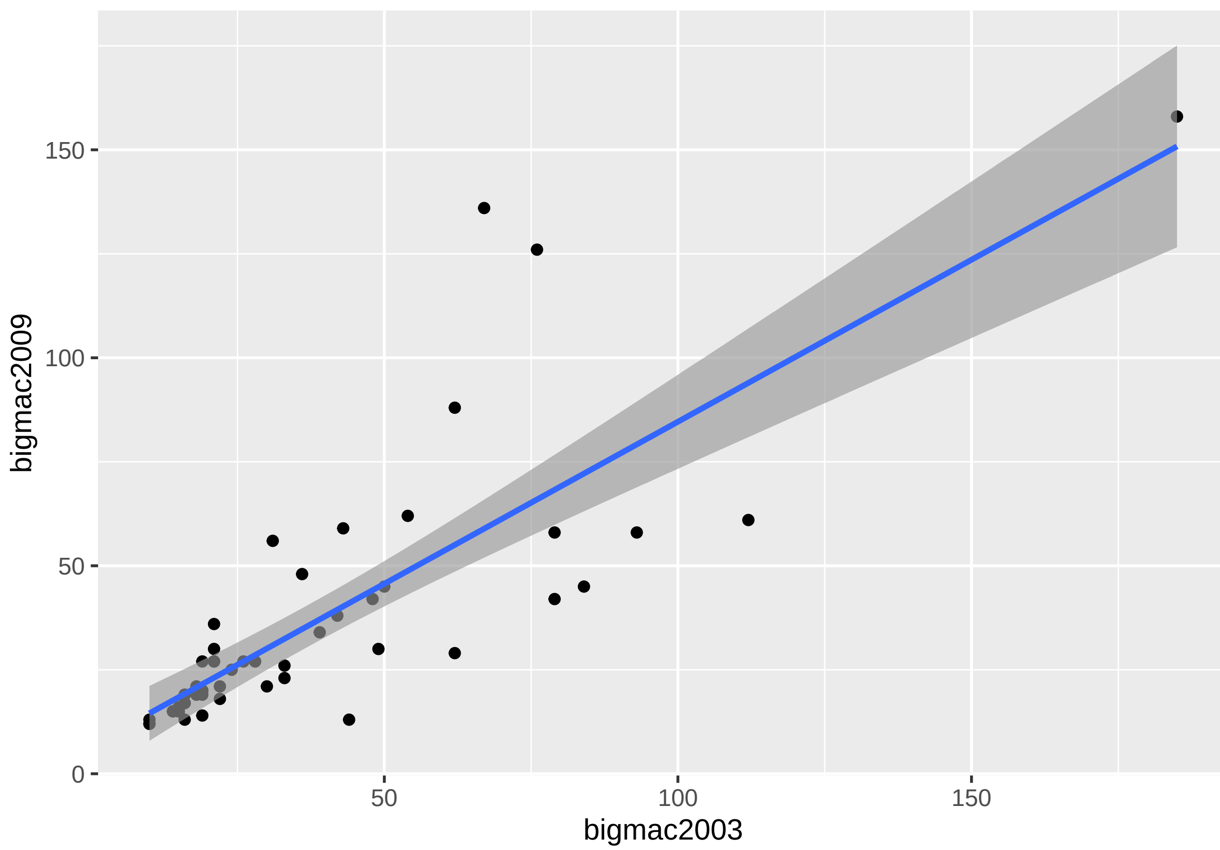
```
##    bigmac2009      bread2009        rice2009       bigmac2003
##  Min.   : 12.00   Min.   : 8.00   Min.   : 8.00   Min.   : 10.00
##  1st Qu.: 17.25   1st Qu.:13.00   1st Qu.:11.00   1st Qu.: 16.50
##  Median : 25.50   Median :19.00   Median :17.00   Median : 22.00
```

```
##  Mean   : 35.35   Mean   :23.02   Mean   :22.34   Mean   : 36.74
##  3rd Qu.: 42.00   3rd Qu.:25.75   3rd Qu.:26.50   3rd Qu.: 47.00
##  Max.   :158.00   Max.   :84.00   Max.   :74.50   Max.   :185.00
##    bread2003        rice2003
##  Min.   : 6.0   Min.   : 5.00
##  1st Qu.:12.0   1st Qu.:12.00
##  Median :18.0   Median :16.00
##  Mean   :21.5   Mean   :19.46
##  3rd Qu.:25.0   3rd Qu.:22.00
##  Max.   :89.0   Max.   :96.00
```

```
plot10<-ggplot(UBSprices,aes(x=bigmac2003, y=bigmac2009))+
  geom_point()+
  stat_smooth(method="lm")+
  geom_smooth(method='lm', formula= y~x)

plot10
```

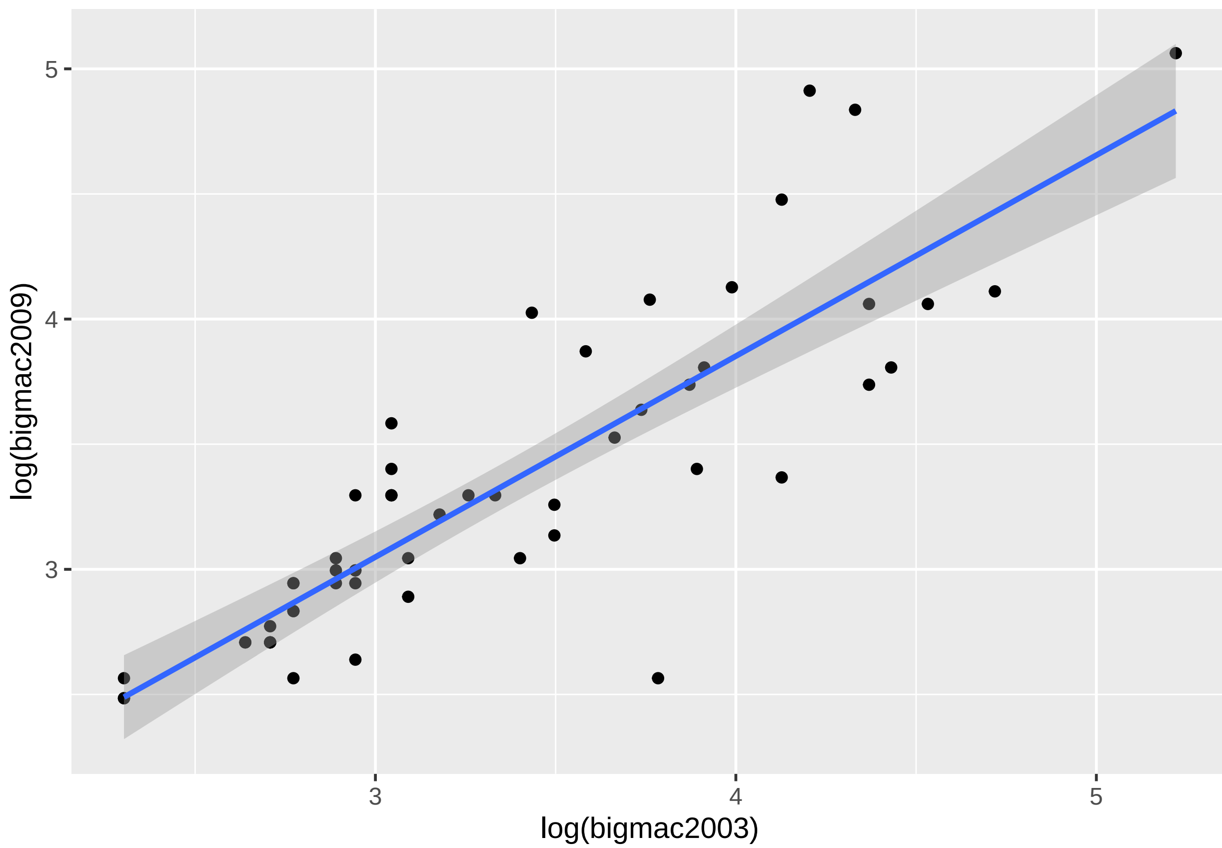
```
## `geom_smooth()` using formula 'y ~ x'
```



2.4.2 The simple linear regression will not be a best way to fit a model to this distribution because most of the data is clustered is one region. In both the axes it has a very skewed distribution with having very less values at the higher end of axes. Due to skewed distribution , it is not a good idea.

```r
plot102<-ggplot(UBSprices,aes(x=log(bigmac2003), y=log(bigmac2009)))+
  geom_point()+
  stat_smooth(method="lm")
plot102
```

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



After the log transformation , it linearly distributes the points across the axes somewhat uniformly. After the transformation there also visible a simple linear relation among both the variables. This is attributed to normal distribution around the axes after log transform. Hence it makes more sense to run simple linear regression for this log transformed model.

Answer 11

```r
summary(ftcollinssnow)
```

```
##       YR1           Early            Late      
##  Min.   :1900   Min.   : 0.50   Min.   : 4.50  
##  1st Qu.:1923   1st Qu.: 9.20   1st Qu.:21.60  
##  Median :1946   Median :14.20   Median :32.00  
##  Mean   :1946   Mean   :16.74   Mean   :32.04  
##  3rd Qu.:1969   3rd Qu.:21.80   3rd Qu.:41.40  
##  Max.   :1992   Max.   :54.90   Max.   :60.30  
```

```r
colnames(ftcollinssnow)
```

```
## [1] "YR1"   "Early" "Late"
```

```r
m11<-lm(Early~Late,ftcollinssnow)
summary(m11)
```

```
##
## Call:
## lm(formula = Early ~ Late, data = ftcollinssnow)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.469  -7.194  -2.868   6.025  35.304
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.67951    2.84831   4.452 2.41e-05 ***
## Late         0.12685    0.08169   1.553    0.124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.83 on 91 degrees of freedom
## Multiple R-squared:  0.02581,    Adjusted R-squared:  0.01511
## F-statistic: 2.411 on 1 and 91 DF,  p-value: 0.1239
```

As per the t test for slope for the linear model , we get the p value of 0.124, which is somewhat lower

Answer 3

2.9.1 $E(Y|X=u) = \beta_0 + \beta_1 u$

$E(Y|Z=u) = \gamma_0 + \gamma_1 z$

$Z = aX + b$

$E(Y|Z=z) = \gamma_0 + \gamma_1(au+b)$

$E(Y|Z=z) = \gamma_0 + \gamma_1 b + \gamma_1 a u$

comparing

$\beta_0 = \gamma_0 + \gamma_1 b$

$\beta_1 = \gamma_1 a$

$$\boxed{\gamma_1 = \beta_1/a \qquad \gamma_0 = \beta_0 - \beta_1 b/a}$$

$\hat{\sigma}^2 = \frac{RSS}{n-2}$ $\quad$ $\mathrm{var}(\beta_1|X) = \hat{\sigma}^2 \frac{1}{SXX}$

$\mathrm{var}(\gamma_1/Z) = \hat{\sigma}^2 \cdot \frac{1}{SZZ}$

This will remain constant because the response variable hasn't changed, hence the residual sum of square and df will remain constant

$S_{xx} = \sum (x_i - \bar{x})^2$

$S_{zz} = \sum (z_i - \bar{z})^2 = \sum (a x_i + b - a\bar{x} - b)^2$

$S_{zz} = a^2 S_{xx}$

$\mathrm{Var}(\beta_1 z_i \mid z) = \frac{\mathrm{var}(\beta_1 \mid x)}{a^2} = \sigma^2 \frac{1}{a^2 S_{xx}}$

Similarly for

$\mathrm{Var}(\beta_0 \mid x) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$

$\mathrm{Var}(Y_0 \mid x) = \sigma^2 \left(\frac{1}{n} + \frac{(ax+b)^2}{a^2 S_{xx}} \right)$

$= \sigma^2 \left(\frac{1}{n} + \frac{x}{S_{xx}} + \frac{2axb}{a S_{xx}} + \frac{b^2}{a^2 S_{xx}} \right)$

2.9.2 $E(y \mid x) = \beta_0 + \beta_1 x$

$d E(y \mid x) = d\beta_0 + d\beta_1 x$

$E(dy \mid x) = d\beta_0 + \beta_1 x$

$E(dy \mid x) = d\beta_0 + d\beta_1 x$

$\delta_0 = d\beta_0 \qquad \delta_1 = d\beta_1$

estimate of variance remains ~~const~~ constant because n predictor doesn't change

The t-test for slope
i.e

$$\beta_1 = \frac{\sigma^2}{\sqrt{Sxx}}$$

2.9.1 $Szz = a^2 Sxx$

so $\beta_1 = \frac{\sigma^2}{\sqrt{Sxx}}$, $\gamma_1 = \frac{S^2}{\sqrt{a^2 Sxx}}$

$= \frac{\sigma^2}{a\sqrt{Sxx}}$

$= a\beta_1$

so t test for slope for β_1 & γ_1 are in a ratio of $\underline{a}$

2.9.2 t test for both the β_1 & δ_1 will remain constant because Sxx will be constant & S^2 will be constant