STAT 51200 HW #1

1.

Length means by age

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 98.342 | 124.847 | 152.563 | 193.800 | 221.720 | 252.597 | 269.868 | 306.25 |

*tapply(Length, Age, mean)*

Length variances by age

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 808.231 | 697.286 | 411.667 | 867.457 | 985.696 | 1105.080 | 869.382 | 1802.916 |

*tapply(Length, Age, var)*

Scale means by age

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 2.386 | 3.132 | 4.078 | 6.209 | 8.105 | 7.700 | 8.517 | 10.198 |

*tapply(Scale, Age, mean)*

Scale variances by Age

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| .716 | .736 | .719 | 2.329 | 2.405 | 2.405 | 3.013 | 1.349 |

*tapply(Scale, Age, var)*

Graph of Length vs. Age

Chart, scatter chart

Description automatically generated

*plot(tapply(avgs$Length, avgs$Age, mean), xlab='Age',ylab='Average Length')*

*abline(lm(avgs$Length~avgs$Age))*

(Continued on next page)

Graph of Length standard deviation vs. Age

Chart, scatter chart

Description automatically generated

*plot(tapply(Length,Age, sd), ylab='Standard Deviation', xlab='Age')*

*tapply(Length,Age, sd)*

Table of SD vs. Age

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 28.429 | 26.406 | 20.289 | 29.452 | 31.395 | 33.242 | 29.485 | 42.460 |

*tapply(Length,Age, sd)*

Summary of Information

Min. 1st Qu. Median Mean 3rd Qu. Max.

20.29 27.92 29.47 30.15 31.86 42.46

*summary(tapply(Length,Age, sd))*

2.

- Quality, clarity, and helpfulness are closely correlated to each other with an average score of around .96. However, easiness and raterinterest are not very closely correlated both with each other and the three dimensions listed previously. The correlation of easiness and raterinterest with the other dimensions and their correlations are predominantly below .5.

*my\_data <- Rateprof[, c(8,9,10, 11, 12)]*

*cor(my\_data)*

3.

2.2.1. All cities with points below the line Y=X have rice prices that have decreased by a certain amount in 2009 compared to prices in 2003 with respect to the amount of minutes that a worker in this area has to work to buy the item. Cities with points above the line have increased rice prices in 2009 compared to prices in 2003.

2.2.2

Vilnius had the largest increase in price from 2003 to 2009.

Mumbai had the largest decrease in price from 2003 to 2009.

*UBSprices$difference <- (rice2009 - rice2003)*

*sorted <- UBSprices[order(UBSprices$difference, decreasing = TRUE),]*

*sorted*

2.2.4

1. Not every location has an increase in price from 2003 to 2009.

2. There are three dimensions to the data but the linear regression line only takes into account two of them.

4.

2.4.1

Manila, Caracas, Sofia, and Jakarta are unusual because their prices have increased from 2003 to 2009 as the ratio of 2009 price to 2003 price is higher. These four mentioned are the cities with the most extreme increases from 2003 to 2009. This is unexpected because one would expect that prices would stay relatively the same because the way price is measured accounts for inflation. Nairobi is also an extreme case because it has the highest time to work to buy a big mac by a long shot for both 2003 and 2009.

*plot(bigmac2003, bigmac2009)*

*abline(0,1)*

*abline(lm(bigmac2003~bigmac2009), col='red', lty=2)*

*legend(115, 50, legend = c("Y=X","2009 vs. 2003"), col = c('black','red'),*

*lty=1:2, cex = .8)*

*sorted*

*row.names(UBSprices)*

*text(bigmac2003, bigmac2009, labels = row.names(UBSprices), cex=.7, pos = 3)*

2.4.2

1. It is likely not a good idea to use simple linear regression because the line does not really account for the change in price for each city.

2.4.3 Additionally, prices of big macs in 2003 and 2009 likely have little to no influence on each other so it doesn’t make much sense to establish a relationship between the two. Although the price in 2009 is likely to be slightly dependent on the 2003 price, it’s not correlated significantly.

2.4.4

This plot looks much better because it’s much easier to visualize the outliers in the data when they’re reduced to the same scale as the smaller inputs. It also helps to account for the starting prices in each city. Instead of the difference being measured in direct numbers, it’s being measured in percentage changes.

*plot(log(bigmac2003), log(bigmac2009))*

*abline(0,1)*

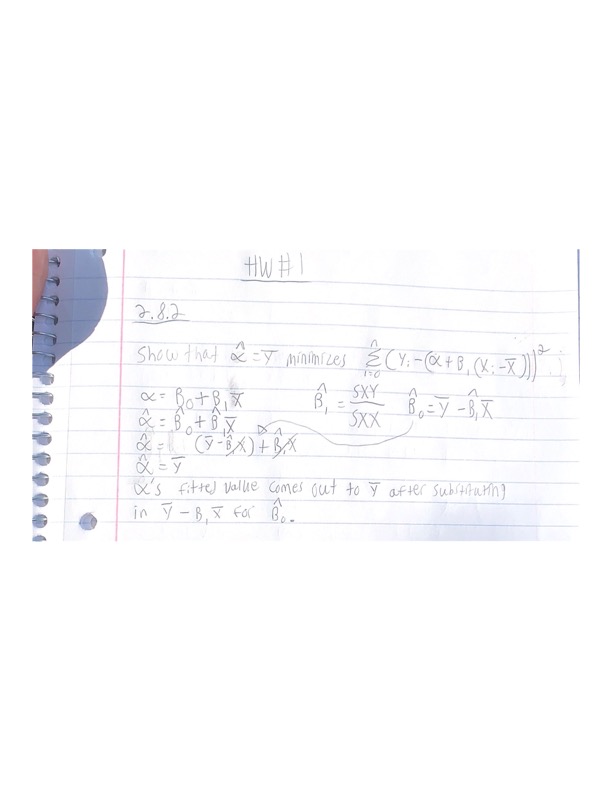
*abline(lm(log(bigmac2003)~log(bigmac2009)), col='red', lty=2)*

2.8

2.8.1

α is the intercept plus the slope \* average of x. It is essentially the average value of Y as you’re taking the slope multiplied by the average x value. However, since the B1 \* xbar term is also subtracted later on in the equation, it mostly represents the intercept still.

2.8.2



This result makes a lot of sense because the Bo + B1xbar is essentially just the average Y-value as stated in 2.8.1 because you take the slope of the linear regression line, multiply it by the average X value and add the intercept to it.

2.8.3

Var(α) = E(Ybar^2) - (E(Ybar))^2

Cov(α) =