Homework 4

Animesh Sengupta

10/4/2022

```
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)
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

Answer 4.2

```
a=(Transact$t1+Transact$t2)/2
d=(Transact$t1-Transact$t2)
head(Transact,5)
```

```
##    t1   t2  time
## 1   0 1166  2396
## 2   0 1656  2348
## 3   0  899  2403
## 4 516 3315 13518
## 5 623 3969 13437
```

```
colnames(Transact)
```

```
## [1] "t1"   "t2"   "time"
```

```
m1=lm(time~t1+t2,data=Transact)
m2=lm(time~a+d,data=Transact)
m3=lm(time~t2+d,data=Transact)
m4=lm(time~t1+t2+a+d,data=Transact)
summary(m1)
```

```
##
## Call:
## lm(formula = time ~ t1 + t2, data = Transact)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.36944  170.54410   0.847    0.398
## t1            5.46206    0.43327  12.607   <2e-16 ***
## t2            2.03455    0.09434  21.567   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = time ~ a + d, data = Transact)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.3694   170.5441   0.847    0.398
## a             7.4966     0.3654  20.514  < 2e-16 ***
## d             1.7138     0.2548   6.726 1.12e-10 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

```
summary(m3)
```

```
##
## Call:
## lm(formula = time ~ t2 + d, data = Transact)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.3694   170.5441   0.847    0.398
## t2            7.4966     0.3654  20.514   <2e-16 ***
## d             5.4621     0.4333  12.607   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

```
summary(m4)
```

```
##
## Call:
## lm(formula = time ~ t1 + t2 + a + d, data = Transact)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.36944  170.54410   0.847    0.398
## t1            5.46206    0.43327  12.607   <2e-16 ***
## t2            2.03455    0.09434  21.567   <2e-16 ***
## a                  NA         NA      NA       NA
## d                  NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

```
cor(Transact$t2,d)
```

```
## [1] -0.986424
```

```
cor(Transact$t1,Transact$t2)
```

```
## [1] 0.7715669
```

4.2.1

As per the model m4 summary, the coefficient estimates are NA. This is because of the fact that the predictors a and b used are linear combination of other predictors t1 and t2. Since it is a linear combination the inverse of matrix to find the estimators gets not invertable.

4.2.2

As per summary of all the 4 models, the intercept and the rest of statistical measures like Residual standard error, Multiple R squared, F-statistics are same. The estimators of t1 and t2 in m1,m4 and m3(for t2) are of same value too. Meanwhile the estimator of "d" changes in all the models it is used in. the estimate for t2 changes from m1 and m3 as well.

4.2.3

We can attribute this difference in estimate due to the correlation between t1 and other predictors. The correlation changes between t2 vs t1-t2 and with t1 vs t2. As you can see from the correlation calculation , it is higher between t2 vs t1-t2 than t2 vs t1.

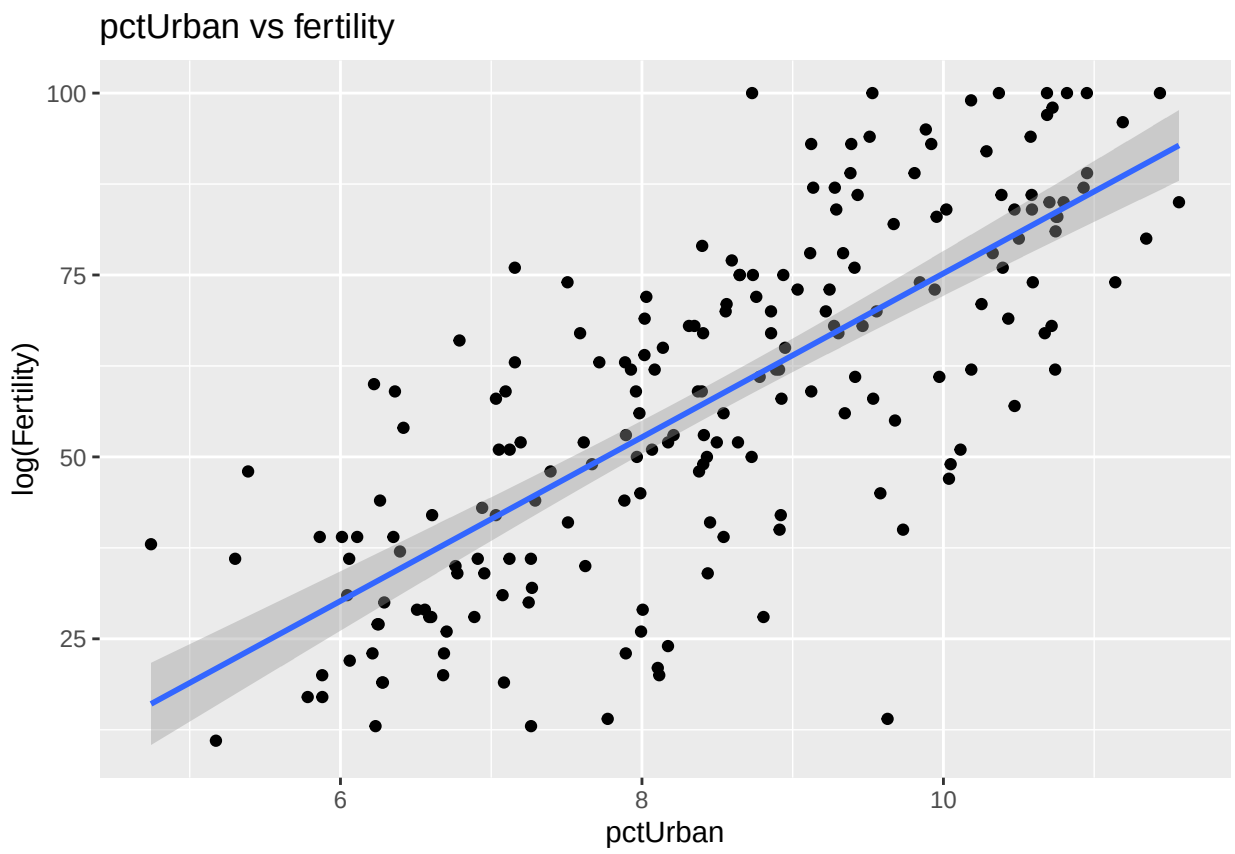
Answer 4.6

```
unml<-lm(log(fertility)~pctUrban,UN11)
summary(unml)
```

```
##
## Call:
## lm(formula = log(fertility) ~ pctUrban, data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87723 -0.31862  0.00894  0.29466  0.73472
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.500963   0.071397  21.023  < 2e-16 ***
## pctUrban    -0.010163   0.001143  -8.892 3.86e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3768 on 197 degrees of freedom
## Multiple R-squared:  0.2864, Adjusted R-squared:  0.2828
## F-statistic: 79.06 on 1 and 197 DF,  p-value: 3.858e-16
```

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

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



The estimated coefficient B0 or the intercept provides the value of the log(fertility) when pctUrban is 0. while the coefficient of pctUrban or b1 provides us the interpretation that with every unit increase in pctUrban the log(fertility) decreases by 0.01

4.7

```
colnames(UN11)
```

```
## [1] "region"    "group"     "fertility" "ppgdp"     "lifeExpF"  "pctUrban"
```

```
unml2<-lm(log(fertility)~log(ppgdp)+lifeExpF,UN11)
summary(unml2)
```

```
##
## Call:
## lm(formula = log(fertility) ~ log(ppgdp) + lifeExpF, data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61778 -0.16891  0.03731  0.17591  0.61072
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.50736    0.12707  27.601  < 2e-16 ***
## log(ppgdp)  -0.06544    0.01781  -3.675 0.000307 ***
## lifeExpF    -0.02824    0.00274 -10.306  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.248 on 196 degrees of freedom
## Multiple R-squared:  0.6926, Adjusted R-squared:  0.6894
## F-statistic: 220.8 on 2 and 196 DF,  p-value: < 2.2e-16
```

```
fert_change=log(0.25*exp(0.065))
fert_change
```

```
## [1] -1.321294
```

As per the emodel fitting, the estimate for log(ppgdp) came out to be -0.065. With a 25% increase in ppgdp , we calculated the log of exp of estimate with 0.25 to verify the decrease in fertility by approx 1.4%

Answer 4.9

4.9.1

The intercept explains the salary in $ for males and that is 24697 since male is assigned 0. The slope of 3340 is the part of salary that females earn less than their male counterparts.

4.9.2

Due to inlcusion of an entirely new estimator i.e year in the model the prediction estimate of salary changes and so does the estimate of sex as well. It means that the year explains more variability in salary i.e more increase in experience correlates to higher salary. Interestingly due to experience in year, it aslo values highly experiences female candidates than their male counterparts. This means that with year as the estimate in the salary prediction model, the weight of sex positively affects the salary.

Answer 4.13

4.13.1

```
colnames(MinnWater)
```

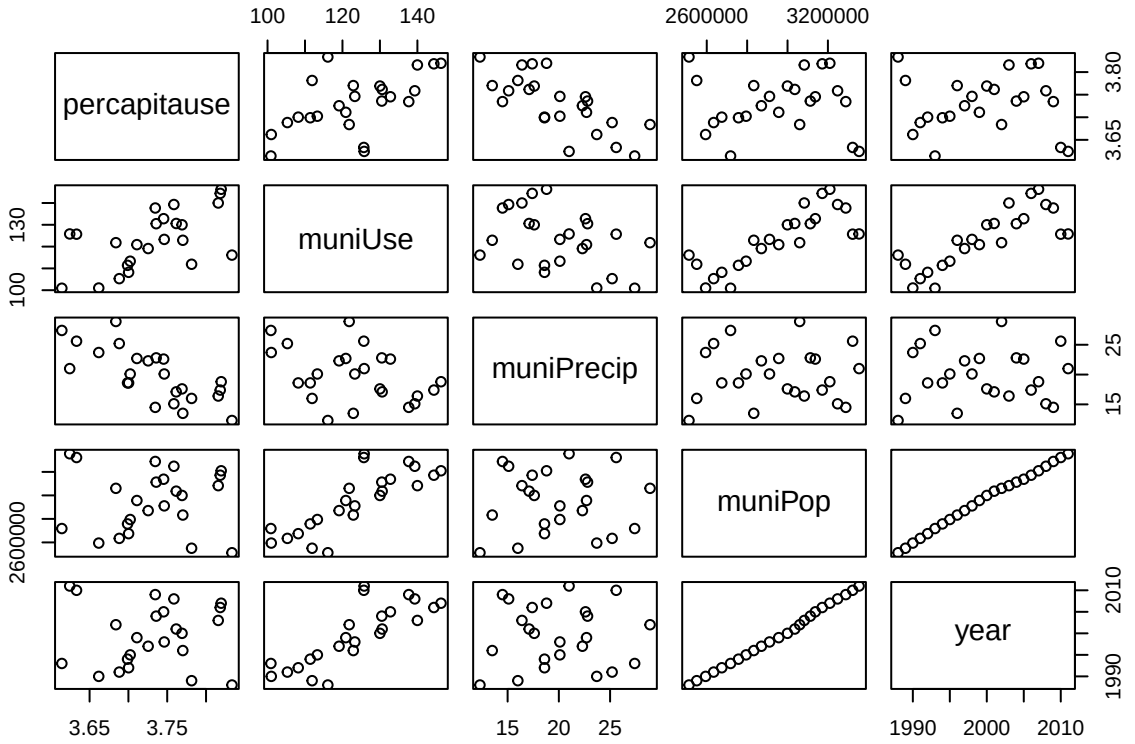
```
## [1] "year"       "allUse"     "muniUse"    "irrUse"     "agPrecip"
## [6] "muniPrecip" "statePop"   "muniPop"
```

```
percapitause=log(10^6*MinnWater$muniUse/MinnWater$muniPop)

mwlm<-lm(percapitause~year+muniPrecip+log(muniPop),data = MinnWater)
summary(mwlm)
```

```
##
## Call:
## lm(formula = percapitause ~ year + muniPrecip + log(muniPop),
##     data = MinnWater)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.087690 -0.032781  0.000155  0.034694  0.080204
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.537117  11.508965   1.089    0.289
## year         -0.011132   0.014141  -0.787    0.440
## muniPrecip   -0.010559   0.002135  -4.946 7.78e-05 ***
## log(muniPop)  0.917355   1.138236   0.806    0.430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04434 on 20 degrees of freedom
## Multiple R-squared:  0.5503, Adjusted R-squared:  0.4828
## F-statistic: 8.157 on 3 and 20 DF,  p-value: 0.0009628
```

```
mwcorr<-pairs(~percapitause+muniUse+muniPrecip+muniPop+year,data = MinnWater)
```



As per the fitted model with new response variable of log of perCapitaUse, there were a few significant changes in the estimates. changing the response variable changed the intercept quite significantly as expected. This can be attributed to the difference in scales of percapitause and muniUse. Although the general trend and effects of the estimates remained similar. Both year and muniPrecip still negatively effects the response while log(muniPop) shows a high effect. As per the scatterplot matrix too, there is not high correlation modelling between percapitause and muniUse as well. There maybe a weak model fit although.