Homework 4

Rohan Sadale

18 Feb 2016

```
library(alr4)
```

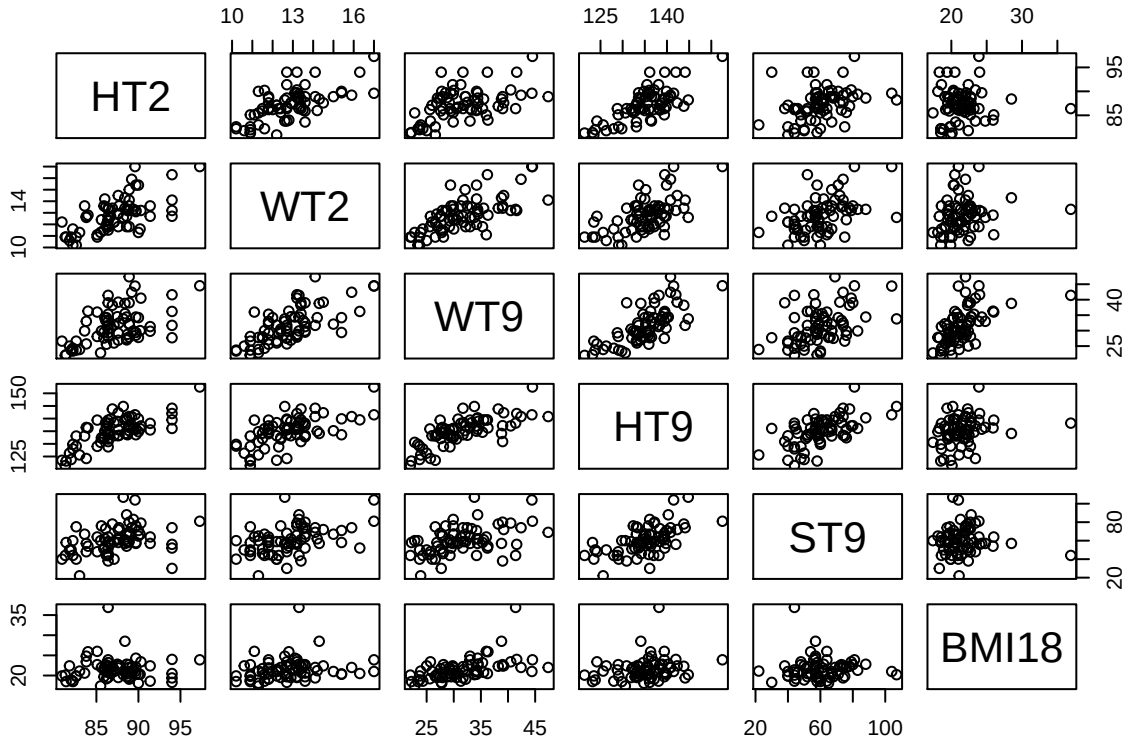
```
## Warning: package 'alr4' was built under R version 3.2.3
```

```
## Warning: package 'car' was built under R version 3.2.3
```

```
## Warning: package 'effects' was built under R version 3.2.3
```

3.3

```
onlyGirls <- subset(BGSall, BGSall$Sex==1)
plot(~ HT2 + WT2 + WT9 + HT9 + ST9 + BMI18, onlyGirls)
```



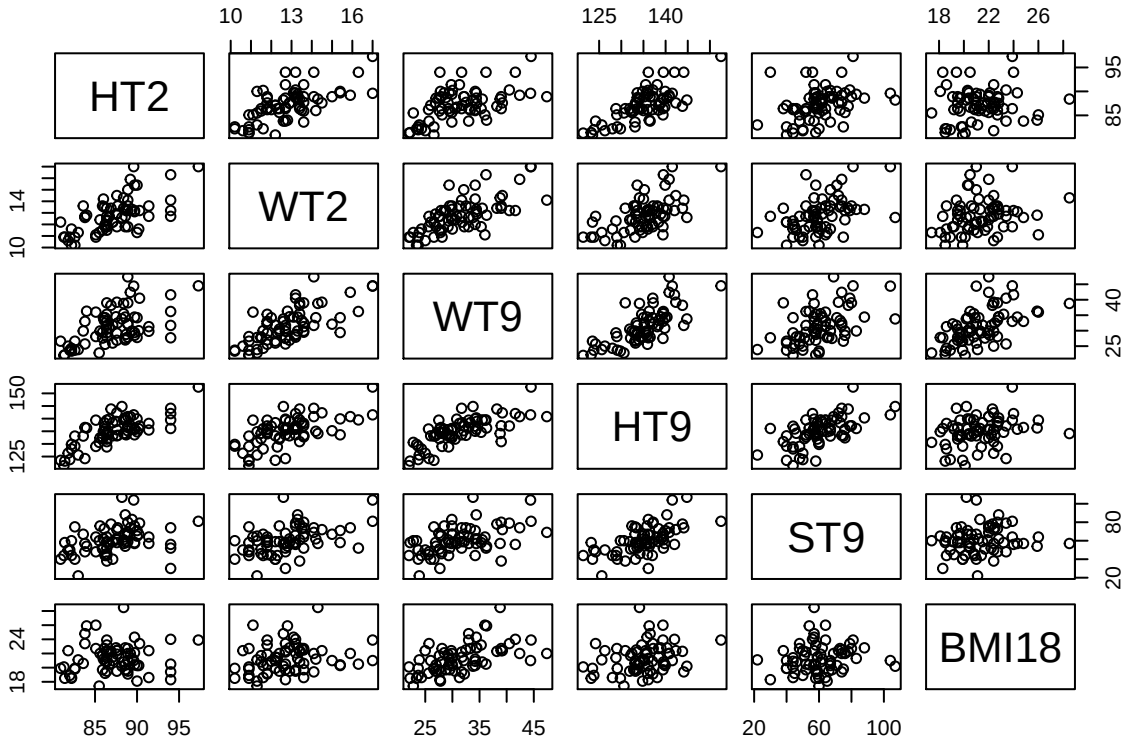
3.3.1

From the plot we can observe that HT2, WT2, WT9, HT9 and ST9 are correlated with each other and we

can fit a simple linear regression line between these pairs. However this is not the case with BMI18. BMI18 doesn't seem to be correlated with any predictors except a very weak correlation with WT9.

We can see that the graph is a bit difficult to interpret because the points appear to be clustered. This is due to girl with BMI above 35. So let's us remove the data point and see if we can improve our inference from scatterplot.

```
onlyGirls1 <- subset(onlyGirls, onlyGirls$BMI18 < 35)
plot(~ HT2 + WT2 + WT9 + HT9 + ST9 + BMI18, onlyGirls1)
```



Now we can see correlation between WT9 and BMI18, but the correlation between BMI18 and rest other predictors is almost similar.

```
cor(onlyGirls[, c("HT2", "WT2", "HT9", "WT9", "ST9", "BMI18")])
```

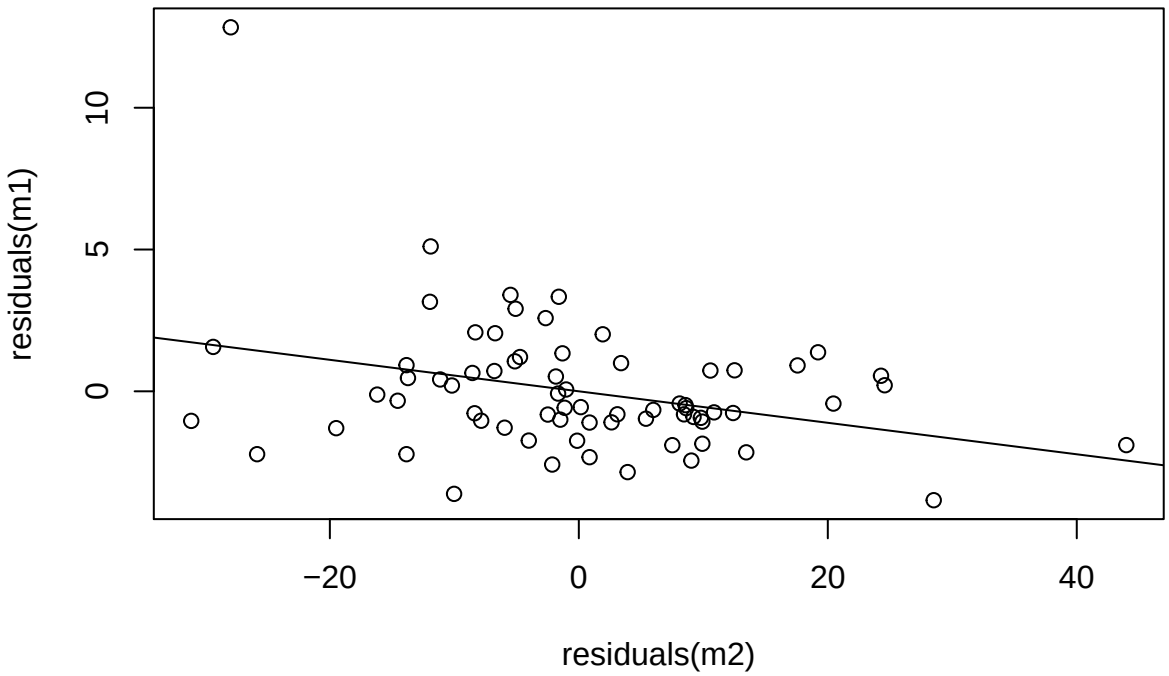
```
##              HT2       WT2       HT9       WT9         ST9       BMI18
## HT2   1.00000000 0.6445495 0.7383562 0.5229277 0.361724146 0.042573733
## WT2   0.64454954 1.0000000 0.6071247 0.6925390 0.451581158 0.190947873
## HT9   0.73835617 0.6071247 1.0000000 0.7276123 0.603368147 0.236907969
## WT9   0.52292768 0.6925390 0.7276123 1.0000000 0.453004062 0.545925753
## ST9   0.36172415 0.4515812 0.6033681 0.4530041 1.000000000 0.005603061
## BMI18 0.04257373 0.1909479 0.2369080 0.5459258 0.005603061 1.000000000
```

```
#Without girl with BMI > 35
cor(onlyGirls1[, c("HT2", "WT2", "HT9", "WT9", "ST9", "BMI18")])
```

```
##              HT2       WT2       HT9       WT9       ST9      BMI18
## HT2   1.00000000 0.6465875 0.7424247 0.5408305 0.3609554 0.08611157
## WT2   0.64658751 1.0000000 0.6063157 0.6996999 0.4611681 0.22285437
## HT9   0.74242471 0.6063157 1.0000000 0.7309602 0.6189165 0.26073658
## WT9   0.54083055 0.6996999 0.7309602 1.0000000 0.4944499 0.56540342
## ST9   0.36095536 0.4611681 0.6189165 0.4944499 1.0000000 0.12933644
## BMI18 0.08611157 0.2228544 0.2607366 0.5654034 0.1293364 1.00000000
```

From the correlation matrix we can see that most of the information summarized above for the plot holds true. Also from the last column of the matrix we can see very weak correlation of predictors with BMI18. Only WT9 has a better correlation with BMI18.

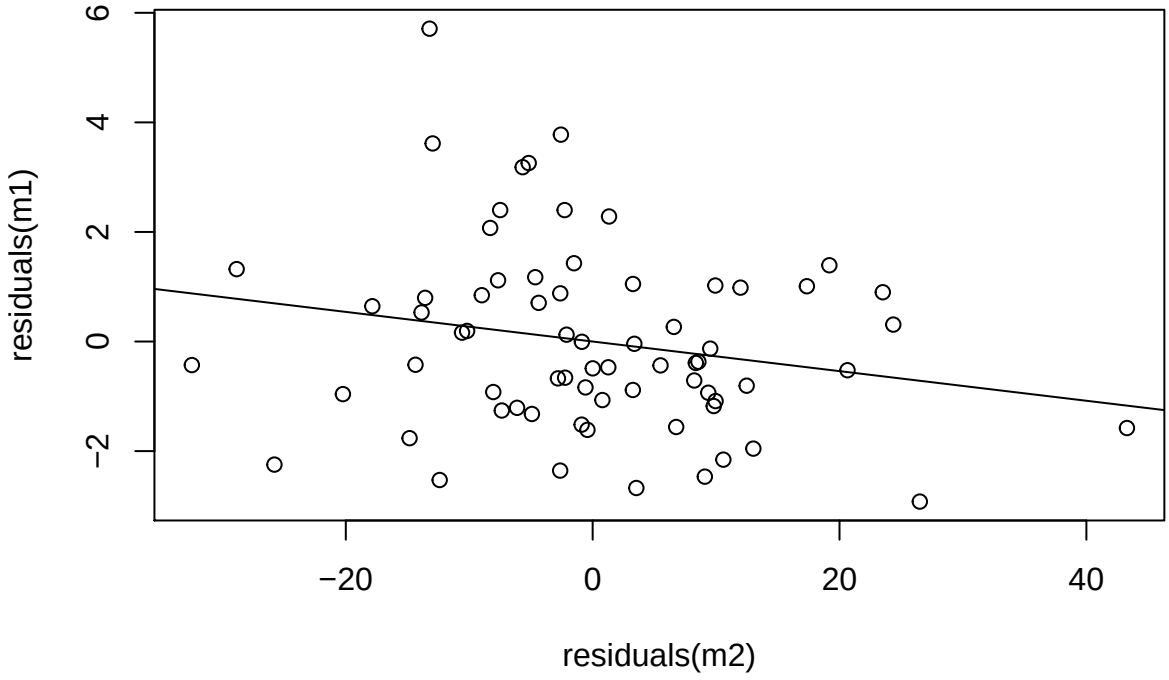
```
m1 <- lm(BMI18 ~ WT9, onlyGirls)
m2 <- lm(ST9 ~ WT9, onlyGirls)
m3 <- lm(residuals(m1)~residuals(m2))
plot(residuals(m1)~residuals(m2))
abline(m3)
```



3.3.2

The added variable plot for ST9 after WT9 shows that after adjustment BMI18 and ST9 are negatively related. This may be due to the presence of the outlier which creates the negative correlation. So if we delete the outlier, we get below plot -

```r
onlyGirls1 <- subset(onlyGirls, onlyGirls$BMI18 < 35)
m1 <- lm(BMI18 ~ WT9, onlyGirls1)
m2 <- lm(ST9 ~ WT9, onlyGirls1)
m3 <- lm(residuals(m1)~residuals(m2))
plot(residuals(m1)~residuals(m2))
abline(m3)
```



Though there is still negative correlation, but it's weak as compared to previous plot.

```r
m1 <- lm(BMI18 ~ HT2 + WT2 + WT9 + HT9 + ST9, onlyGirls)
summary(m1)
```

3.3.3

```
##
## Call:
## lm(formula = BMI18 ~ HT2 + WT2 + WT9 + HT9 + ST9, data = onlyGirls)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0948 -1.2186 -0.2533  1.0090 10.4951
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.855335   8.781156   3.514 0.000817 ***
## HT2         -0.193997   0.130819  -1.483 0.142996
## WT2         -0.317779   0.278736  -1.140 0.258505
## WT9          0.419762   0.075211   5.581  5.2e-07 ***
## HT9          0.008057   0.096344   0.084 0.933613
## ST9         -0.044416   0.022219  -1.999 0.049853 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.14 on 64 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.3996
## F-statistic: 10.19 on 5 and 64 DF,  p-value: 3.294e-07
```

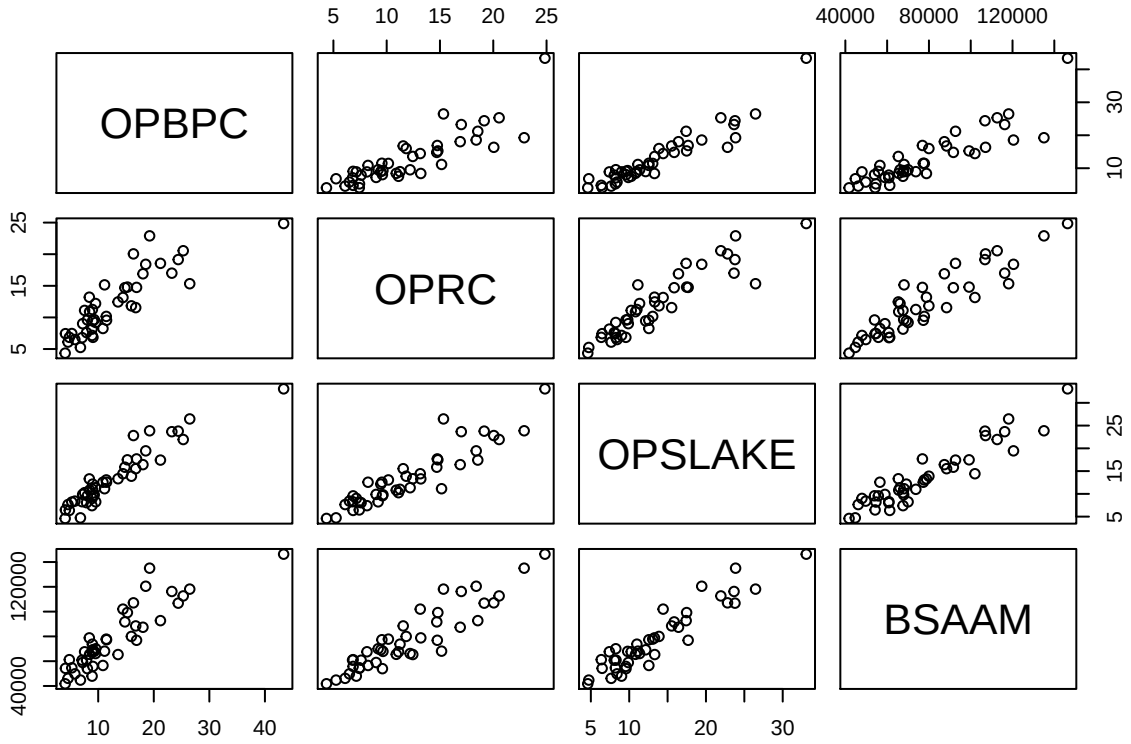
From the summary we can see R-squared is 0.443 which says that the regression explains 44.3% of variation in BMI.

The hypothesis tested by the t-values are that each of the slope values(HT2, WT2, WT9, HT9, ST9) = 0 with other slope values arbitary vs. slope value != 0 with other slope values arbitrary.

Also we can see that only WT9 and ST9 have p-value below 0.05 which says that these are the only variables that can explain the variation in BMI.

3.6

```
plot(~OPBPC + OPRC + OPSLAKE + BSAAM, water)
```



3.6.1

```
cor(water[,c("OPBPC", "OPRC", "OPSLAKE", "BSAAM")])
```

```
##             OPBPC      OPRC   OPSLAKE     BSAAM
## OPBPC   1.0000000 0.8647073 0.9433474 0.8857478
## OPRC    0.8647073 1.0000000 0.9191447 0.9196270
## OPSLAKE 0.9433474 0.9191447 1.0000000 0.9384360
## BSAAM   0.8857478 0.9196270 0.9384360 1.0000000
```

From the plot, we can see all the predictors - OPBPC, OPRC, OPSLAKE are strongly correlated with each other and also strongly correlated with the response variable - BSAAM. We can see that OPBPC has stongest correlation with OPSLAKE and lowest with OPRC. OPRC has strongest correlation with BSAAM and lowest correlation with OPBPC. OPBPC has strongest correlation with OPBC and lowest with OPRC. Of all the predictors, we can see that OPSLAKE seems to be strongly correlated with other variables.

The same inference we can make from the correlation matrix.

```
m1 <- lm(BSAAM ~ OPBPC + OPRC + OPSLAKE, water)
summary(m1)
```

3.6.2

```
##
## Call:
## lm(formula = BSAAM ~ OPBPC + OPRC + OPSLAKE, data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15964.1  -6491.8   -404.4   4741.9  19921.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22991.85    3545.32   6.485  1.1e-07 ***
## OPBPC          40.61     502.40   0.081  0.93599
## OPRC         1867.46     647.04   2.886  0.00633 **
## OPSLAKE      2353.96     771.71   3.050  0.00410 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8304 on 39 degrees of freedom
## Multiple R-squared:  0.9017, Adjusted R-squared:  0.8941
## F-statistic: 119.2 on 3 and 39 DF,  p-value: < 2.2e-16
```

The hypothesis tested by the t-values are that each of the slope values(OPBPC, OPRC, OPSLAKE) = 0 with other slope values arbitary vs. slope value != 0 with other slope values arbitrary.From the p-values we can see that OPRC and OPSLAKE have significant level less that 0.05 which indicates that if we add them to the regression model we can better explain the variation in response BSAAM.

```
m1 <- lm(BMI18 ~ WT2 + WT9 + WT18, BGSgirls)
summary(m1)
```

4.1

```
##
## Call:
## lm(formula = BMI18 ~ WT2 + WT9 + WT18, data = BGSgirls)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1037 -0.7432 -0.1240  0.8320  4.3485
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.30978    1.65517   5.020 4.16e-06 ***
## WT2         -0.38663    0.15145  -2.553    0.013 *
## WT9          0.03141    0.04937   0.636    0.527
## WT18         0.28745    0.02603  11.044  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.333 on 66 degrees of freedom
## Multiple R-squared:  0.7772, Adjusted R-squared:  0.767
## F-statistic: 76.73 on 3 and 66 DF,  p-value: < 2.2e-16
```

```r
ave <- (BGSgirls$WT2 + BGSgirls$WT9 + BGSgirls$WT18)/3
lin <- BGSgirls$WT18 - BGSgirls$WT2
quad <- BGSgirls$WT2 - 2*BGSgirls$WT9 + BGSgirls$WT18

m2 <- lm(BMI18 ~ ave + lin + quad, BGSgirls)
summary(m2)
```

```
##
## Call:
## lm(formula = BMI18 ~ ave + lin + quad, data = BGSgirls)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1037 -0.7432 -0.1240  0.8320  4.3485
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.30978    1.65517   5.020 4.16e-06 ***
## ave         -0.06778    0.12751  -0.532    0.597
## lin          0.33704    0.07466   4.514 2.68e-05 ***
## quad        -0.02700    0.03976  -0.679    0.499
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.333 on 66 degrees of freedom
## Multiple R-squared:  0.7772, Adjusted R-squared:  0.767
## F-statistic: 76.73 on 3 and 66 DF,  p-value: < 2.2e-16
```

From the plot we can see -

- The coefficient of determination is same in both models.
- All residuals are identical
- Intercepts are same
- From the first model we can see that BMI depends on WT2 and WT18. In the second mode we can see that BMI depends on lin i.e. linear combination of WT18 and WT2. The interpretation in transformed scale seems easier as only linear time is significant in explaining BMI. Thus, we can describe the change in BMI over time as increasing by the same amount each year.

```r
a <- (Transact$t1 + Transact$t2)/2
d <- (Transact$t1 - Transact$t2)

m1 <- lm(time ~ t1+t2, Transact)
m2 <- lm(time ~ a+d, Transact)
m3 <- lm(time~ t2+d, Transact)
m4 <- lm(time~ t1 + t2 + a + d, Transact)

# Model 1
summary(m1)
```

4.2

```
##
## 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
```

```
# Model 2
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
```

```
# Model 3
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
```

```r
# Model 4
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
```

4.2.1

- While fitting model m4, the variability in time is explained by t1 and t2. Thus, while building model, when a and d are encountered, they aren't able to provide any additional information in explaining time. This is because a and d are linear combination of t1 and t2. Thus, a and d are marked as NA in the summary table.

4.2.2 From the summary statistics we can see -

- The coefficient of determination is same in all models.
- All residuals are identical
- Intercepts are same
- Degrees of freedom are same

- Slopes are different.
- t-values and p-values are different
- Standard errors are different

4.2.3

- The coefficient of t2 is different in models 1 and 3 because, in model 1 it represents change in time keeping t1 as constant. On the other hand, in model 3, it represents change in time keeping d as constant. As d = t1 - t2, it kind of puts a restriction on how t2 can change which also cause the response to changes.