Takehome-final

Dandong Tu

On my honor, I have not had any form of communication about this exam with any other individual(including other students, teaching assistants, instructors, etc.)–Dandong Tu

1

```
b1=powerTransform(cbind(heterogeneity,mobility)~region,data=Angell)
```

Based on the summary of **b1**(see detailed summary in Appendix), it suggests that we should use **log** transformation for both **heterogeneity** and **mobility**.

```
m1=lm(moralIntegration~log(heterogeneity)+log(mobility)+region,data=Angell)
```

```
inverseResponsePlot(m1)
boxCox(m1)
```

The inverseResponsePlot and the boxCox shows that the lambda=1 seems appropriate for the response.

```
summary(m1)
```

After making the scatterplotmatrix and based on the summary of **m1**,it seems log(**mobility**) and **region** is somehow statisticlly significant(with p-value close to 0.1) while log(**heterogeneity**) is statisticlly significant.

```
m1.a=lm(moralIntegration~log(heterogeneity),data=Angell)
m1.b=lm(moralIntegration~log(heterogeneity)+region,data=Angell)
m1.c=lm(moralIntegration~log(heterogeneity)+log(mobility),data=Angell)
```

1st **anova(m1.a,m1)** 2nd **anova(m1.b,m1)** 3rd **anova(m1.c,m1)** (see Appendix for details)

The first anova test shows that we reject the reduced model **m1.a**. Second test shows that the reduced model is somehow adequate with **p-value=0.096**(close to 0.1). The third anova test shows that we have weak evidence against the null hypothesis that the reduced model is adequate with large **p-value=0.3599**. Therefore, we chose model **m1.c** that contains log(**heterogeneity**) and log(**mobility**)

```
m2=lm(moralIntegration~poly(log(heterogeneity),2)+poly(log(mobility),2),data=Angell)
```

```
summary(m2)
```

The summary of model **m2** shows that the 1st degree of polynomial is statisticlly significant for both log(**heterogeneity**) and log(**mobility**), while 2nd degree of polynomial for both log(**heterogeneity**) and log(**mobility**) are not statisticlly significant, so that we do not include second degree of polynomial.

```
m3=lm(moralIntegration~log(heterogeneity)*log(mobility),data=Angell)
Anova(m3)
```

The Type II anova test for model **m3** shows that the interation is not statisticlly significant. So that we use the model without the interation.

```
ncvTest(m1.c)
```

Based on the ncvTest, we observed a high **p-value**(**0.9779**). Thus, we do not have enough evidence to reject the null(the Variance is constant), and we assume the variance is constant.

2

```
summary(m1.c)
```

```
coef(m1.c)
```

```
##        (Intercept) log(heterogeneity)      log(mobility)
##          42.203611          -3.783073          -5.729764
```

We interpretation for the log(**heterogeneity**) coefficient estimate is the following: if we increase the log(**heterogeneity**)by one unit,keeping the log(**mobility**) fixed, the **moralIntegration** index would decrease by -3.783 units on average.

The interpretation for the log(**mobility**) coefficient estimate is the following: if we increase log(**mobility**) by one unit, keeping the log(**heterogeneity**) fixed, the **moralIntegration** index would decrease by -5.73 units on average.

```
(rp=residualPlots(m1.c))
```

Based on the result of residualPlots tests, all plots look like null plots. And none of the tests has small significance levels, providing no evidence against the mean function.(detailed graphs and test show in supporting material)

At this point, we assume our model is appropriate. And no further changes is needed at this point.

3

```
influenceIndexPlot(m1.c,id.n=4)
```

The Cook's distance determined that the observation **Rochester**, **SanDiego**, **Portlandoregon** and **Houston** are 4 most influentials.Please note the value of Cook's distance shows that **Louisville** and **Tulsa** are similar influential with **Portlandoregon** and **Houston**

The Bonferroni p-value are all close to 1, which indicates that there is no outliers.

```
Angell2=Angell[-c(1,18,29,31),] #Rochester,Sandiego,portlandoregon and Houston
m1.d=lm(moralIntegration~log(heterogeneity)+log(mobility),data=Angell2)
compareCoefs(m1.c,m1.d)
```

```
##
## Call:
## 1: lm(formula = moralIntegration ~ log(heterogeneity) + log(mobility),
##   data = Angell)
## 2: lm(formula = moralIntegration ~ log(heterogeneity) + log(mobility),
##   data = Angell2)
##                    Est. 1   SE 1 Est. 2   SE 2
## (Intercept)        42.204  3.458 42.353  3.261
## log(heterogeneity) -3.783  0.540 -3.878  0.491
## log(mobility)      -5.730  0.870 -5.729  0.841
```

The result shows that the coefficient estimates do not have much difference between the the model with selected most 4 influential points and the model without selected most 4 influential points.

Appendix

1

```
summary(b1)
```

```
## bcPower Transformations to Multinormality
##                Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## heterogeneity   -0.2340           0      -0.6689       0.2008
## mobility        -0.4034           0      -1.0687       0.2619
##
## Likelihood ratio tests about transformation parameters
##                                LRT df         pval
## LR test, lambda = (0 0)  2.266886  2 3.219230e-01
## LR test, lambda = (1 1) 41.510852  2 9.683516e-10
```

```
testTransform(b1,c(0,0))
```

```
##                               LRT df     pval
## LR test, lambda = (0 0) 2.266886  2 0.321923
```

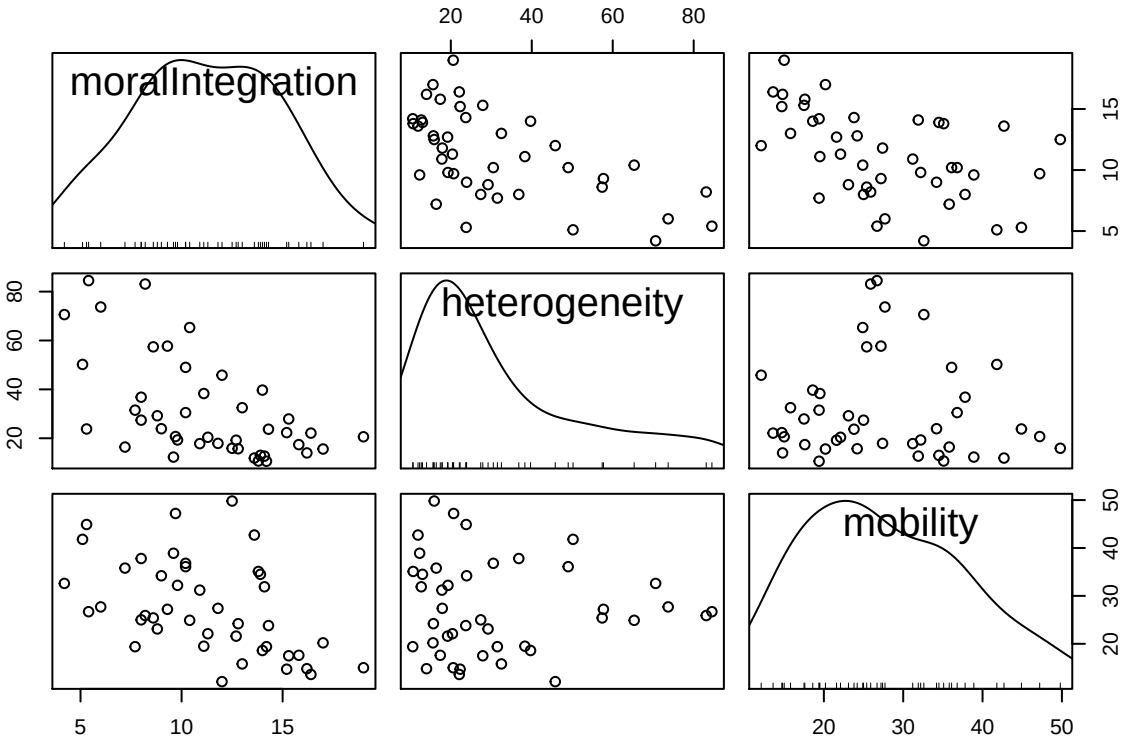
```
testTransform(b1,c(1,0))
```

```
##                               LRT df         pval
## LR test, lambda = (1 0) 30.99393  2 1.861028e-07
```

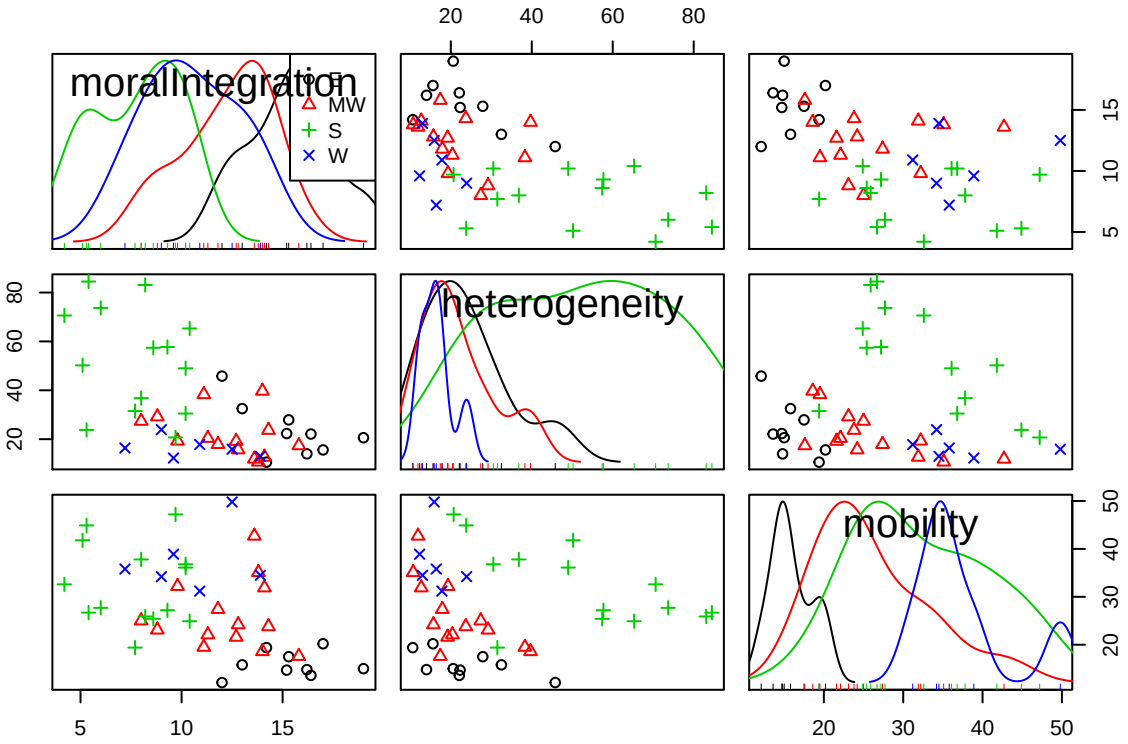
```
testTransform(b1,c(0,1))
```

```
##                               LRT df         pval
## LR test, lambda = (0 1) 14.83196  2 0.0006015614
```

```
scatterplotMatrix(~moralIntegration+heterogeneity+mobility,data = Angell,reg.line=FALSE,smooth=FALSE)
```

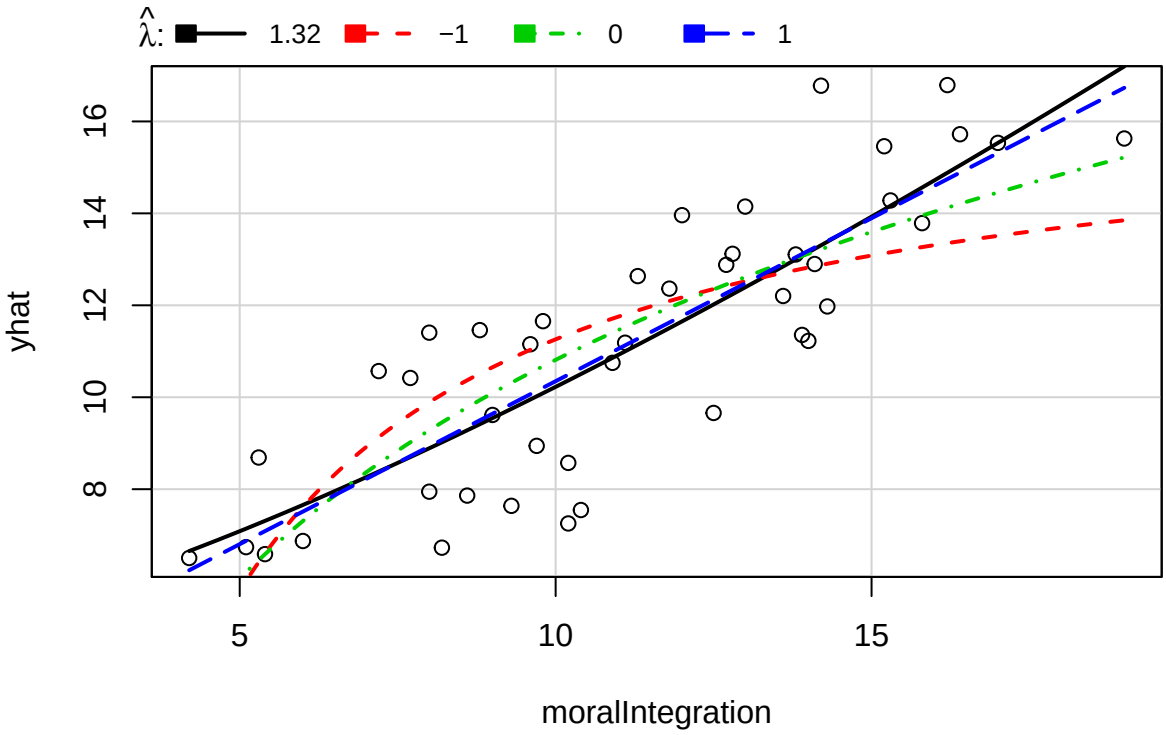


```
scatterplotMatrix(~moralIntegration+heterogeneity+mobility|region,data = Angell,reg.line=FALSE,smooth=F
```



It seems there is no specific patterns in scatterplot matrix.

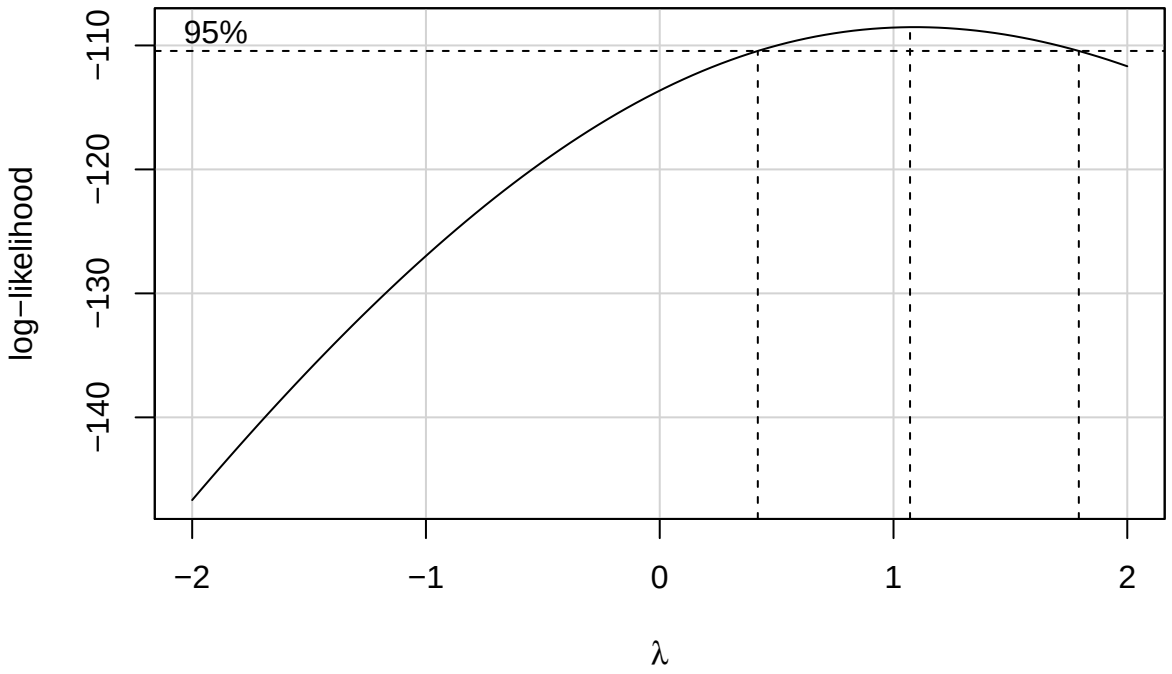
```
inverseResponsePlot(m1)
```




```
##       lambda      RSS
## 1  1.317358 109.5181
```

```
## 2 -1.000000 165.7426
## 3  0.000000 128.2721
## 4  1.000000 110.5400
```

```
boxCox(m1)
```




```
summary(m1)
```

```
##
## Call:
## lm(formula = moralIntegration ~ log(heterogeneity) + log(mobility) +
##     region, data = Angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4052 -1.4439 -0.0865  1.4695  3.3714
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         32.6188     7.1611   4.555 5.53e-05 ***
## log(heterogeneity)  -2.9023     0.9856  -2.945  0.00556 **
## log(mobility)       -3.0317     1.7760  -1.707  0.09620 .
## regionMW            -1.8468     1.1901  -1.552  0.12920
## regionS             -3.1968     2.0214  -1.581  0.12229
## regionW             -3.0837     1.7456  -1.767  0.08555 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.053 on 37 degrees of freedom
## Multiple R-squared:  0.709,  Adjusted R-squared:  0.6697
## F-statistic: 18.03 on 5 and 37 DF,  p-value: 4.899e-09
```

```r
anova(m1.a,m1)
```

```
## Analysis of Variance Table
##
## Model 1: moralIntegration ~ log(heterogeneity)
## Model 2: moralIntegration ~ log(heterogeneity) + log(mobility) + region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     41 353.91
## 2     37 155.90  4    198.01 11.749 2.938e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m1.b,m1.c)
```

```
## Analysis of Variance Table
##
## Model 1: moralIntegration ~ log(heterogeneity) + region
## Model 2: moralIntegration ~ log(heterogeneity) + log(mobility)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     38 168.18
## 2     40 169.85 -2   -1.6721 0.1889 0.8286
```

```r
anova(m1.c,m1)
```

```
## Analysis of Variance Table
##
## Model 1: moralIntegration ~ log(heterogeneity) + log(mobility)
## Model 2: moralIntegration ~ log(heterogeneity) + log(mobility) + region
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     40 169.85
## 2     37 155.90  3     13.95 1.1036 0.3599
```

```r
summary(m2)
```

```
##
## Call:
## lm(formula = moralIntegration ~ poly(log(heterogeneity), 2) +
##     poly(log(mobility), 2), data = Angell)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.135 -1.255 -0.108  1.602  3.538
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    11.2000     0.3174  35.284  < 2e-16 ***
## poly(log(heterogeneity), 2)1 -14.0484     2.1332  -6.586 8.99e-08 ***
## poly(log(heterogeneity), 2)2   1.3846     2.1846   0.634     0.53
## poly(log(mobility), 2)1      -13.8827     2.1432  -6.478 1.26e-07 ***
## poly(log(mobility), 2)2        2.1908     2.1748   1.007     0.32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.082 on 38 degrees of freedom
## Multiple R-squared:  0.6927, Adjusted R-squared:  0.6604
## F-statistic: 21.42 on 4 and 38 DF,  p-value: 2.595e-09
```

```
m3=lm(moralIntegration~log(heterogeneity)*log(mobility),data=Angell)
Anova(m3)
```

```
## Anova Table (Type II tests)
##
## Response: moralIntegration
##                                  Sum Sq Df F value    Pr(>F)
## log(heterogeneity)               208.457  1 47.9622 2.713e-08 ***
## log(mobility)                    184.061  1 42.3493 1.023e-07 ***
## log(heterogeneity):log(mobility)   0.346  1  0.0795    0.7794
## Residuals                        169.505 39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ncvTest(m1.c)
```

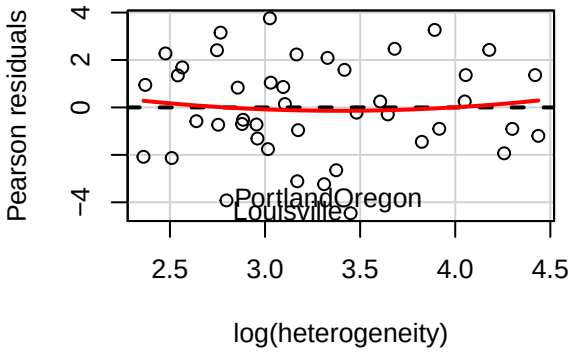
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.0007643731    Df = 1     p = 0.9779435
```

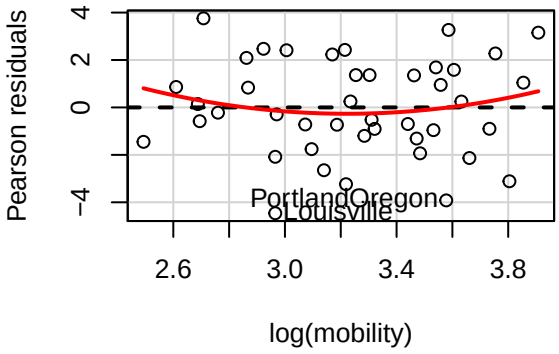
2

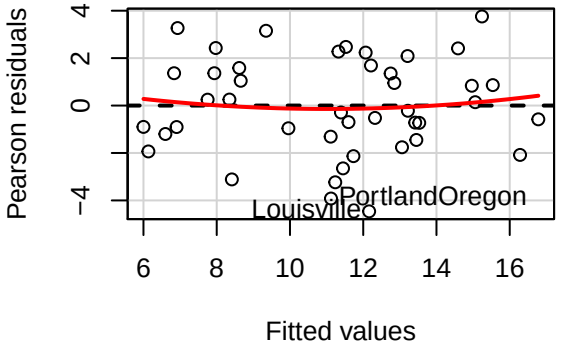
```
summary(m1.c)
```

```
##
## Call:
## lm(formula = moralIntegration ~ log(heterogeneity) + log(mobility),
##     data = Angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4617 -1.2552 -0.2196  1.4745  3.7578
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         42.2036     3.4578  12.205 4.58e-15 ***
## log(heterogeneity)  -3.7831     0.5399  -7.007 1.84e-08 ***
## log(mobility)       -5.7298     0.8703  -6.584 7.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.061 on 40 degrees of freedom
## Multiple R-squared:  0.683,  Adjusted R-squared:  0.6672
## F-statistic: 43.09 on 2 and 40 DF,  p-value: 1.049e-10
```

```
(rp=residualPlots(m1.c, id.n=2))
```







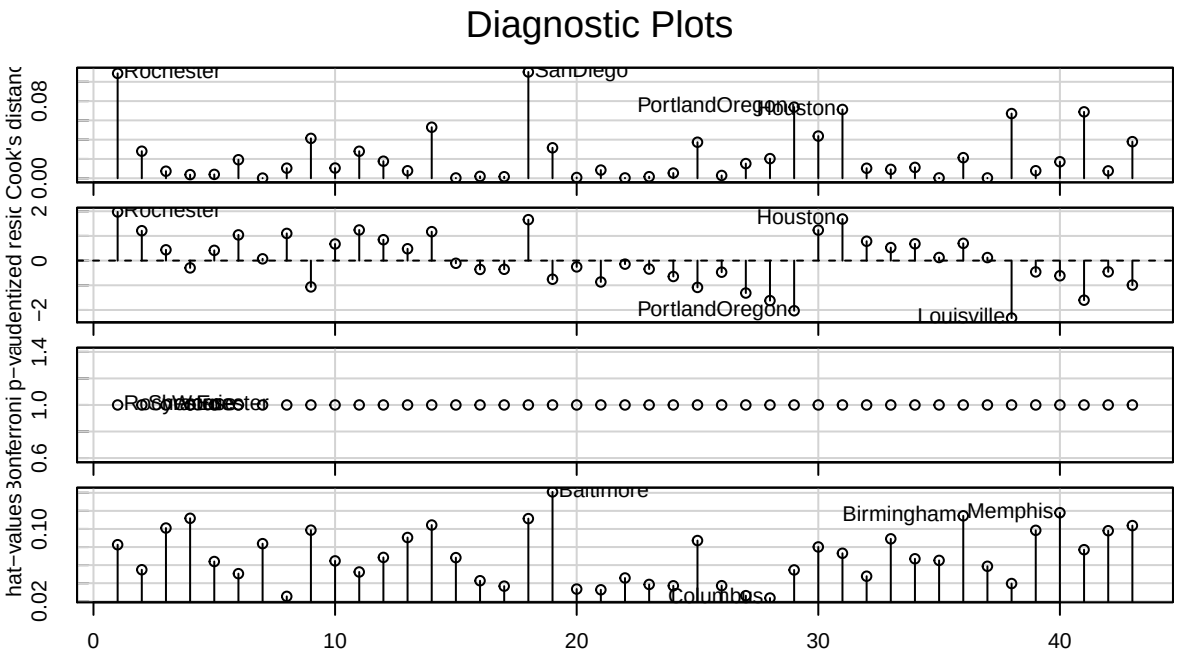

```
##                     Test stat Pr(>|t|)
## log(heterogeneity)      0.433    0.668
## log(mobility)           0.902    0.373
## Tukey test              0.494    0.622
```

3

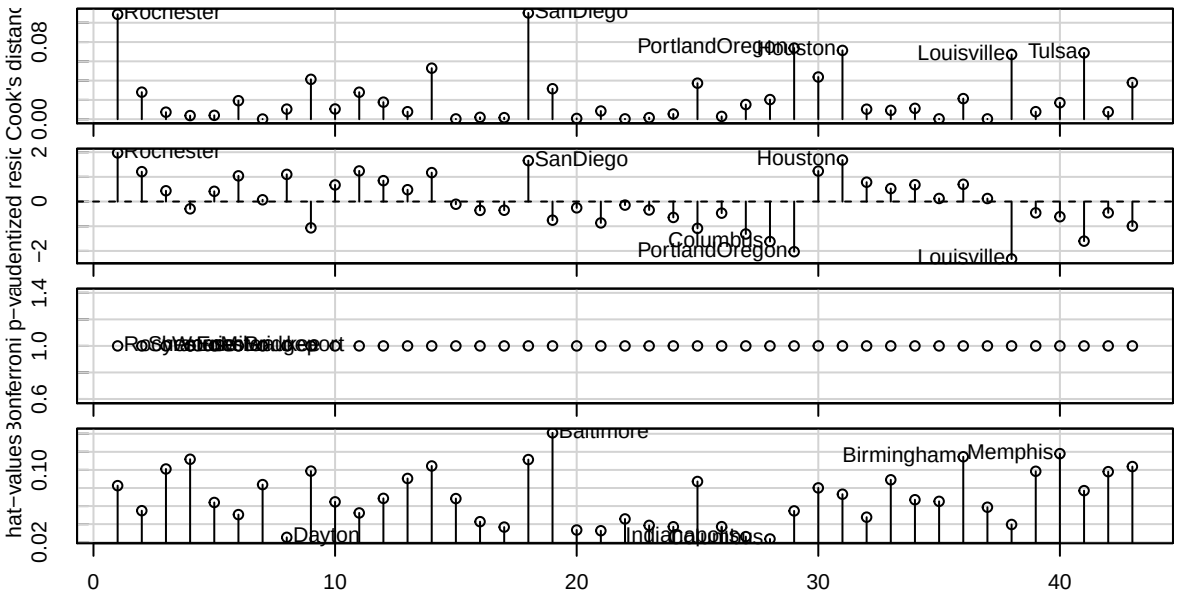
```
outlierTest(m1.c)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##             rstudent unadjusted p-value Bonferonni p
## Louisville -2.328294           0.025174           NA
```

```
influenceIndexPlot(m1.c,id.n=4)
```




```
influenceIndexPlot(m1.c,id.n=6)
```




```
summary(m1.d)
```

```
##
## Call:
## lm(formula = moralIntegration ~ log(heterogeneity) + log(mobility),
##     data = Angell2)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2851 -1.0552 -0.0976  1.5198  2.6721
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         42.3530     3.2610  12.988 3.75e-15 ***
## log(heterogeneity)  -3.8781     0.4911  -7.896 2.27e-09 ***
## log(mobility)       -5.7291     0.8407  -6.815 5.75e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.815 on 36 degrees of freedom
## Multiple R-squared:  0.7398, Adjusted R-squared:  0.7254
## F-statistic: 51.18 on 2 and 36 DF,  p-value: 2.985e-11
```