S631 Final Takehome

Erik Parker

December 12th, 2017

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.) Signed: Erik Parker

Use the data *Angell.txt*, in that data use *moralIntegration* as the response and all other variables as predictors.

1. Using methods and techniques learned in class to determine: Which predictors to include, if polynomials and/or interactions should be included, if weights or other adjustments should be used for the variance, if transformations to predictors and/or the response should be considered.

Based on all of the tests and analysis shown in appendix 1, we see first that *heterogeneity* and *mobility* should both be log transformed when *region* is included alongisde them in a model. When *region* is not included though, log transformation of *mobility* was found to be inappropriate. This was seen visually from analysis of the scatterplot, the summary of the data, and most directly through use of the box-cox method using the *powerTransform* command, and also the *testTransform* command.

Next, F-tests of multiple potential models found that the only predictors to be included in the final model are *heterogeneity* and *region*. When all three regressors are included in the model, we are lead to believe that *region* should be dropped, but we can see from further testing that *region* alone explains the highest portion of the variation seen in the response, and also if *region* is dropped from the model a log transformation of *mobility* is no longer found to be appropriate. Then, if we conclude that the full model is not appropriate due to the insignificance of *region*, and we compare the the resulting, most appropriate, potential two regressor models ($log(heterogeneity) + region$ and $log(heterogeneity) + mobility$) we see that the model with *region* is far and away better than the one with *mobility*.

Furthermore, no interactions or polynomials need to be included in the model, as the interaction between *log(heterogeneity)* and *region* was shown to be non-significant through F-testing, as was the polynomial term for *heterogeneity*. And finally, we also see from the *ncvTest* that no weights or adjustments need to be used for the variance as this test returned a large p-value, meaning we can't reject the null hypothesis that the chosen model has constant variance. The final model then is *moralIntegration ~ log(heterogeneity) + region*.

2. Based on the model obtained so far: interpret at least two coefficient estimates, perform an analysis of residuals and determine if any assumptions made about the mean and variance are appropriate, and make any appropriate changes to the model.

First, an interpretation of the coefficient for *log(heterogeneity)* as shown in appendix 2. This coefficient estimate corresponds change in expected value of the response, *moralIntegration*, for every unit increase in the regressor *log(heterogeneity)*, when all other regressors are kept constant. So, a coefficient of -1.8926 here tells us that for every unit increase in *log(heterogeneity)*, we see an decrease of *moralIntegration* by 1.8926 units. This can also be expressed as an increase in *heterogeneity* by $exp(1) = 2.7183$ units leads to a decrease in *moralIntegration* by 1.8926 units, regardless of the level of the other regressor, *region*.

Next, an interpretation of the estimated coefficient for *regionS* as shown in appendix 2. This estimated coefficient value corresponds to the difference between the sample mean for the baseline level, *regionE*, and this new level, *regionS*, when all other regressors are kept constant. Put another way, we can say that when we move from the Northeast region to the South region, the expected value of our response, *moralIntegration* decreases by 6.11 units on average when we keep the continuous regressor, *log(heterogeneity)* constant.

Finally, an interpretation of the coefficient for *regionMW*. This coefficient estimate corresponds to the change in intercept of the regression line as we move from level *regionE* to *regionMW*, while keeping *log(heterogeneity)* constant. The coefficient of -3.2169 tells us that as we move from the Northeast region to the Midwest one, there is a decrease in *moralIntegration* of 3.2169 units, when we keep *log(heterogeneity)* constant.

Finally, when examining the residuals, it's clear that the plots contained in appendix 2 resemble null-plots with no curvature (so no need to alter the mean function), and no heteroskedasticity (so no need to adjust the variance). So, from this there seems to be no need to change the chosen model yet.

3. With the resulting model from part 2, perform an influence anaylsis by doing the following. First, use Cook's distance to determine which observations are the most influential (up to 4), then determine which observations if any could be considered outliers, finally, compare the coefficient estimates obtained with and without the selected influential observations.

From the *influenceIndexPlot* command shown in appendix 3 we can see that the four most influential observations are Tulsa, Portland (my hometown!), Denver, and Rochester. Though they are the most influential, their values of D_i are still quite small and far below 1, we can say that removal of any of these cases likely won't change our $\hat{\beta}$ estimates much at all, per ALR section 9.5.2.

From the *outlierTest* command, we can see that the largest "outlier", Tulsa, has an insignificant unadjusted p-value and thus an insignificant Bonferonni p-value. This means that there are no observations in this data which could be considered outliers, a conclusion also reflected in the third panel of the diagnostic plot.

Though earlier analysis of the Cook's distance plot revealed that removal of influential observations wasn't really necessary here, it was performed for the sake of this exercise. When we remove the four most influential observations as identified above, we see that the estimated coefficients do change noticeably across the board. In general, the coefficients for the factor levels decrease (become less negative) by roughly 0.5 to 1 units, while the estimates for the intercept (factor level *Northeast*) and the continuous regressor increase (become more negative) by about 0.5 units. So it seems that while no one observation was highly influential, removing four of them at once did enough to change the results of the model noticeably.

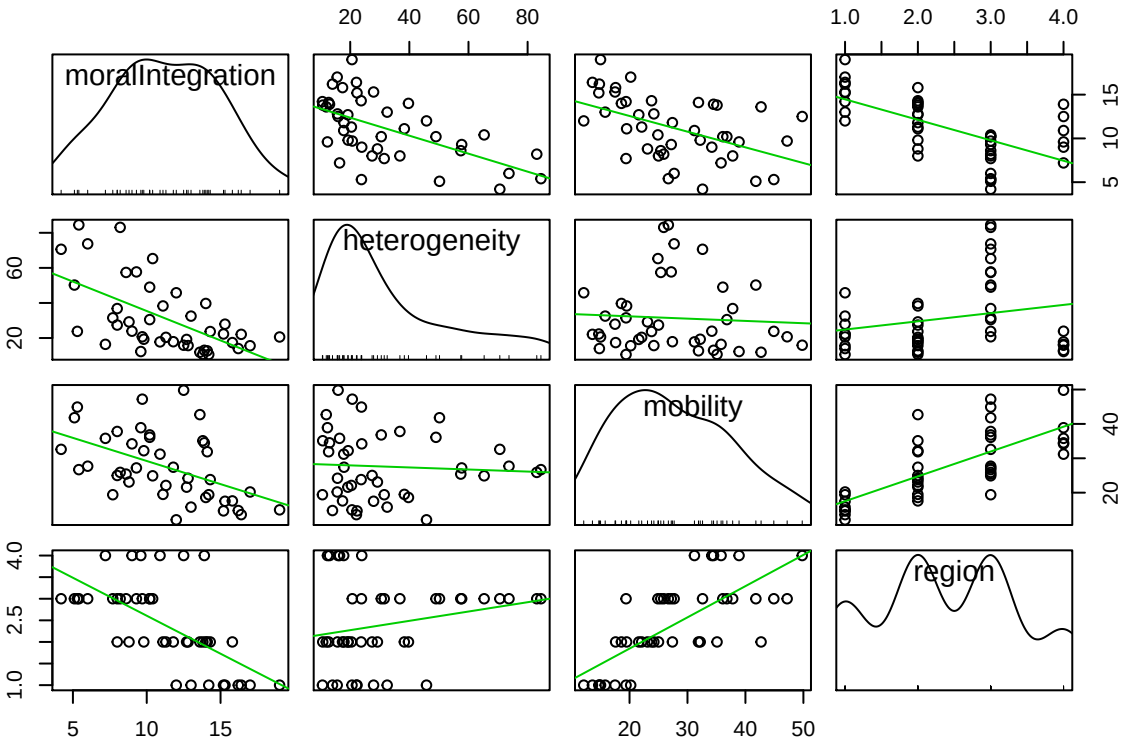
Appendix

1.

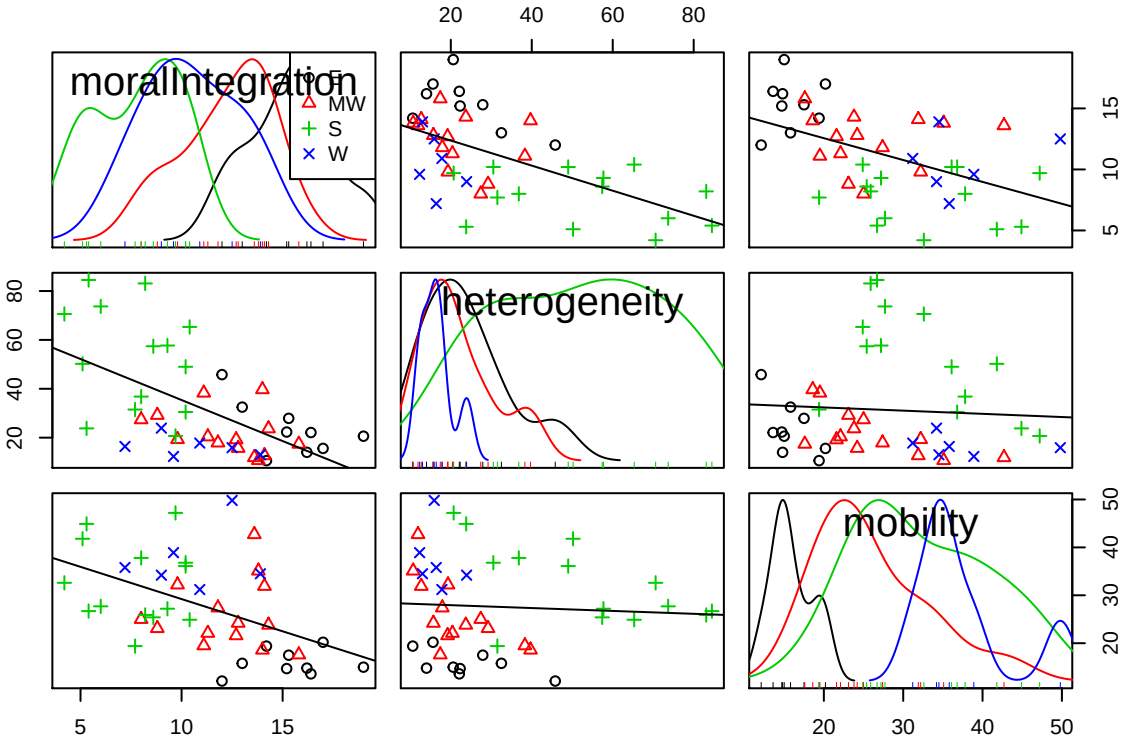
```
rm(list = ls())

library(alr4)

angell <- read.table("Angell.txt")

summary(angell)
```

```
##  moralIntegration heterogeneity      mobility     region
##  Min.   : 4.20    Min.   :10.60   Min.   :12.10   E : 9
##  1st Qu.: 8.70    1st Qu.:16.90   1st Qu.:19.45   MW:14
##  Median :11.10    Median :23.70   Median :25.90   S :14
##  Mean   :11.20    Mean   :31.37   Mean   :27.60   W : 6
##  3rd Qu.:13.95    3rd Qu.:39.00   3rd Qu.:34.80
##  Max.   :19.00    Max.   :84.50   Max.   :49.80
```

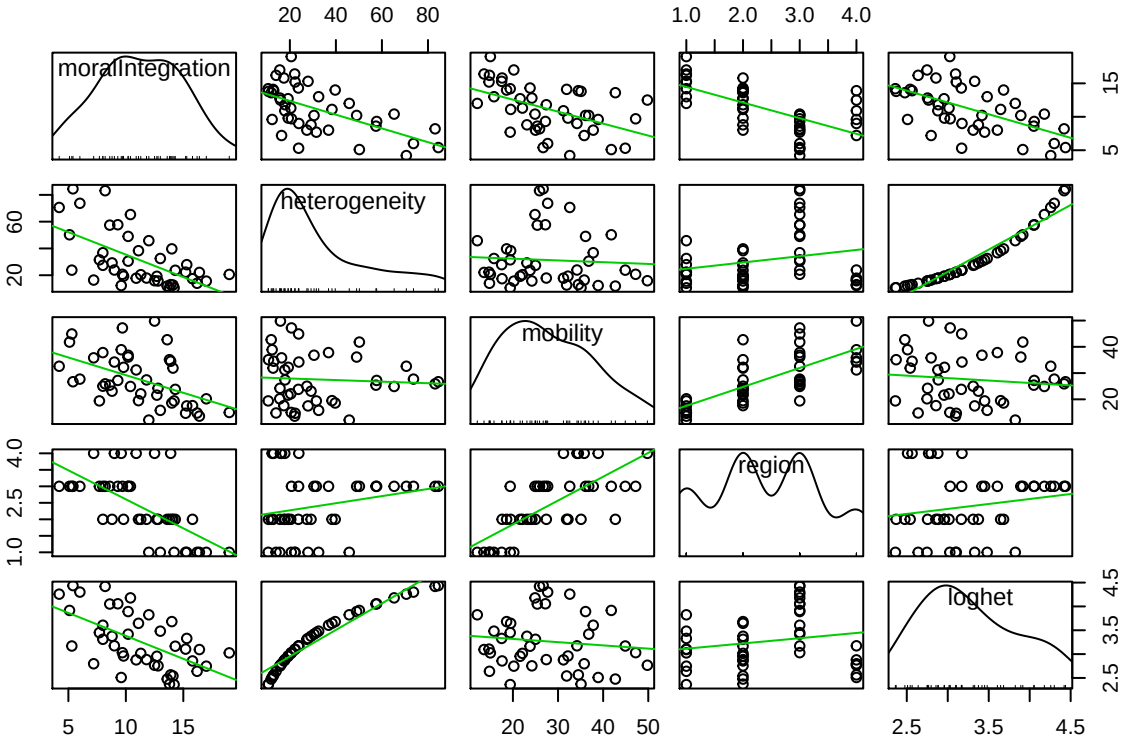
```
scatterplotMatrix(angell, smoother = FALSE)
```




```
# heterogeneity probably needs a transformation, others look okay.
scatterplotMatrix(~moralIntegration + heterogeneity + mobility | region, angell,
    smoother = FALSE)
```




```
angell$loghet <- log(angell$heterogeneity)

scatterplotMatrix(angell, smoother = FALSE)
```




```
# After transformation, relationship of heterogeneity with moralIntegration,
# and distribution looks better.
```

```r
bc1 = powerTransform(cbind(heterogeneity, mobility) ~ 1, angell)
summary(bc1)
```

```
## bcPower Transformations to Multinormality
##               Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## heterogeneity   -0.4232           0      -0.9674       0.1210
## mobility         0.2776           1      -0.5662       1.1214
##
## Likelihood ratio tests about transformation parameters
##                                LRT df         pval
## LR test, lambda = (0 0)  2.759157  2 2.516846e-01
## LR test, lambda = (1 1) 29.270117  2 4.406309e-07
```

```r
# So from this, heterogeneity seems to need a log transformation, mobility
# is rounded to 1, even though it's closer to 0 than it is to 1. So maybe we
# should consider a log transformation of mobility as well.

testTransform(bc1, c(0, 1))
```

```
##                              LRT df       pval
## LR test, lambda = (0 1) 5.135706  2 0.07670003
```

```r
# See a slightly higer p-value when we don't transform mobility, but is it
# different enough to conclude mobility doesn't need a log transformation.

bc2 = powerTransform(cbind(heterogeneity, mobility) ~ region, angell)
summary(bc2)
```

```
## 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
```

```r
# When we consider the factor, region,it is clear that mobility needs to be
# log transformed. From these two tests, it seems like mobility and
# heterogeneity both need to be log transformed, regardless of the inclusion
# of region in the model.


m1a <- lm(moralIntegration ~ log(heterogeneity) + log(mobility) + region, data = angell)
Anova(m1a)
```

```
## Anova Table (Type II tests)
##
## Response: moralIntegration
##                      Sum Sq Df F value   Pr(>F)
## log(heterogeneity)  36.540  1  8.6721 0.005558 **
## log(mobility)       12.278  1  2.9140 0.096195 .
## region              13.950  3  1.1036 0.359930
## Residuals          155.900 37
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# So, maybe region not needed when mobility is log transformed?

m2 <- lm(moralIntegration ~ log(heterogeneity) + region, data = angell)

anova(m2, m1a)
```

```
## Analysis of Variance Table
##
## Model 1: moralIntegration ~ log(heterogeneity) + region
## Model 2: moralIntegration ~ log(heterogeneity) + log(mobility) + region
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     38 168.18
## 2     37 155.90  1    12.278 2.914 0.0962 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# log transformed mobility and heterogeneity improves the model, but we
# still can't reject the null that the model without mobility is good
# enough, BUT we get awfully close.  Maybe model 1a is the best?

m2b <- lm(moralIntegration ~ log(heterogeneity) + log(mobility), data = angell)

anova(m2b, m1a)
```

```
## 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
```

```
# But, this test also tells us that we can't reject the null that the model
# without region is good.

# What about if we don't log transform mobility as potentially suggested by
# the first power transform test?
m2c <- lm(moralIntegration ~ log(heterogeneity) + mobility, data = angell)

anova(m2c, m1a)
```

```
## Analysis of Variance Table
##
## Model 1: moralIntegration ~ log(heterogeneity) + mobility
## Model 2: moralIntegration ~ log(heterogeneity) + log(mobility) + region
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     40 185.96
## 2     37 155.90  3    30.055 2.3777 0.08548 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Also can't reject the reduced model, but it's the closest yet.  Seems to
# be saying that this reduced model is the worst of the three?

Anova(m1a)
```

```
## Anova Table (Type II tests)
##
## Response: moralIntegration
##                      Sum Sq Df F value   Pr(>F)
## log(heterogeneity)  36.540  1  8.6721 0.005558 **
## log(mobility)       12.278  1  2.9140 0.096195 .
## region              13.950  3  1.1036 0.359930
## Residuals          155.900 37
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(m2)
```

```
## Anova Table (Type II tests)
##
## Response: moralIntegration
##                      Sum Sq Df F value    Pr(>F)
## log(heterogeneity)  24.286  1  5.4874   0.02449 *
## region             185.734  3 13.9889 2.694e-06 ***
## Residuals          168.178 38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(m2b)
```

```
## Anova Table (Type II tests)
##
## Response: moralIntegration
##                     Sum Sq Df F value    Pr(>F)
## log(heterogeneity) 208.46  1  49.092 1.836e-08 ***
## log(mobility)      184.06  1  43.347 7.153e-08 ***
## Residuals          169.85 40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(m2c)
```

```
## Anova Table (Type II tests)
##
## Response: moralIntegration
##                     Sum Sq Df F value    Pr(>F)
## log(heterogeneity) 220.72  1  47.479 2.665e-08 ***
## mobility           167.96  1  36.128 4.566e-07 ***
## Residuals          185.96 40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = moralIntegration ~ log(heterogeneity) + region,
##     data = angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7610 -1.3475  0.0564  1.3553  3.5536
```

```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         21.1721     2.5756   8.220 5.87e-10 ***
## log(heterogeneity)  -1.8926     0.8079  -2.343 0.024493 *
## regionMW            -3.2169     0.9005  -3.572 0.000981 ***
## regionS             -6.1121     1.1083  -5.515 2.63e-06 ***
## regionW             -5.3904     1.1325  -4.760 2.80e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.104 on 38 degrees of freedom
## Multiple R-squared:  0.6861, Adjusted R-squared:  0.6531
## F-statistic: 20.77 on 4 and 38 DF,  p-value: 3.851e-09
```

```
summary(m2b)
```

```
##
## 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
```

```
summary(m2c)
```

```
##
## Call:
## lm(formula = moralIntegration ~ log(heterogeneity) + mobility,
##     data = angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4692 -1.3091  0.0089  1.5438  4.2669
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         29.6373     2.1961  13.495  < 2e-16 ***
## log(heterogeneity)  -3.9072     0.5670  -6.891 2.66e-08 ***
## mobility            -0.2056     0.0342  -6.011 4.57e-07 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.156 on 40 degrees of freedom
## Multiple R-squared:  0.653,  Adjusted R-squared:  0.6356
## F-statistic: 37.63 on 2 and 40 DF,  p-value: 6.424e-10
```

```
summary(m1a)
```

```
##
## 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
```

```
# See from this that m1a with log transformed mobility and heterogeneity has
# the highest R^2 value - it explains the most variation in the response of
# the models tested so far, though just barely.  The model which was found
# to be more significant through F-testing, model 2, has an R^2 value just
# below it. Also, we have model 2b, with both log transformed continuous
# regressors, which seems to be a good model - BUT I don't think it is an
# appropriate one, as the log transformation for mobility only seems
# appropriate when region is included in the model. So then we need to look
# at model 2c, and it is substantially worse.  So the only models we need to
# seriously consider are 1a (the full model) and 2 (the reduced one with
# region and log(heterogeneity)). Right now I am leaning towards 2, because
# the levels of region are mostly insignificant in 1a, but I don't think I
# want to drop region. Below I look at the main effects of each regressor in
# a model by itself, to try and determine which one explains the most
# information - and it turns out to be region, so I think that is strong
# evidence that it should be included in the model, leading me to conclude
# that model 2 is the best one.

# Let's try adding an interaction between mobility and region maybe, first
# let's look at R^2 values of models with only those two though, to try and
# determine if they explain similar data.

m3a <- lm(moralIntegration ~ region, data = angell)
```

```r
m3b <- lm(moralIntegration ~ log(mobility), data = angell)
summary(m3a)
```

```
##
## Call:
## lm(formula = moralIntegration ~ region, data = angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2786 -1.6262  0.3833  1.6774  3.6333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.3667     0.7405  20.752  < 2e-16 ***
## regionMW     -3.0881     0.9491  -3.254 0.002356 **
## regionS      -7.6310     0.9491  -8.040 8.37e-10 ***
## regionW      -4.8500     1.1708  -4.142 0.000178 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.221 on 39 degrees of freedom
## Multiple R-squared:  0.6408, Adjusted R-squared:  0.6132
## F-statistic: 23.19 on 3 and 39 DF,  p-value: 8.787e-09
```

```r
summary(m3b)
```

```
##
## Call:
## lm(formula = moralIntegration ~ log(mobility), data = angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7832 -2.4954  0.5508  1.9116  5.0435
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     28.403      4.189   6.780 3.37e-08 ***
## log(mobility)   -5.286      1.279  -4.132 0.000173 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.038 on 41 degrees of freedom
## Multiple R-squared:  0.294,  Adjusted R-squared:  0.2767
## F-statistic: 17.07 on 1 and 41 DF,  p-value: 0.000173
```

```r
# region explains more variation by itself alone, may not need mobility.
m3c <- lm(moralIntegration ~ log(mobility) + region, data = angell)
summary(m3c)
```

```
##
## Call:
## lm(formula = moralIntegration ~ log(mobility) + region, data = angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.2775 -1.6152  0.4015  1.6803  3.6383
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15.0714     4.3543   3.461  0.00134 **
## log(mobility)   0.1072     1.5574   0.069  0.94547
## regionMW       -3.1390     1.2127  -2.588  0.01359 *
## regionS        -7.7054     1.4470  -5.325 4.79e-06 ***
## regionW        -4.9418     1.7843  -2.770  0.00863 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.25 on 38 degrees of freedom
## Multiple R-squared:  0.6408, Adjusted R-squared:  0.603
## F-statistic: 16.95 on 4 and 38 DF,  p-value: 4.678e-08
```

```
# lots of shared explanatory potential, R^2 doesn't increase when both
# present.
m3d <- lm(moralIntegration ~ log(mobility) * region, data = angell)
summary(m3d)
```

```
##
## Call:
## lm(formula = moralIntegration ~ log(mobility) * region, data = angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2813 -1.8219  0.3778  1.5434  3.7740
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.9576    13.6845   0.508    0.614
## log(mobility)             3.0533     4.9608   0.615    0.542
## regionMW                  6.2245    15.9222   0.391    0.698
## regionS                   3.4035    16.1170   0.211    0.834
## regionW                  -9.2934    26.8129  -0.347    0.731
## log(mobility):regionMW   -3.3331     5.5613  -0.599    0.553
## log(mobility):regionS    -3.8146     5.5384  -0.689    0.496
## log(mobility):regionW     0.5069     8.0830   0.063    0.950
##
## Residual standard error: 2.319 on 35 degrees of freedom
## Multiple R-squared:  0.6488, Adjusted R-squared:  0.5786
## F-statistic: 9.237 on 7 and 35 DF,  p-value: 1.986e-06
```

```
Anova(m3d)
```

```
## Anova Table (Type II tests)
##
## Response: moralIntegration
##                       Sum Sq Df F value    Pr(>F)
## log(mobility)          0.024  1  0.0045    0.9471
## region               185.867  3 11.5237 2.096e-05 ***
## log(mobility):region   4.266  3  0.2645    0.8505
## Residuals            188.174 35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# interaction not significant.

Anova(m3c)
```

```
## Anova Table (Type II tests)
##
## Response: moralIntegration
##                Sum Sq Df F value    Pr(>F)
## log(mobility)   0.024  1  0.0047    0.9455
## region        185.867  3 12.2340 9.536e-06 ***
## Residuals     192.440 38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# mobility not needed, even when log transformed.

# So, seems like we might need both log(heterogeneity) and region, but not
# mobility.  Seems like mobility explains the same information as region, so
# it is only significant really when added to a model not already containing
# region.

m3d <- lm(moralIntegration ~ log(heterogeneity), data = angell)
summary(m3d)
```

```
##
## Call:
## lm(formula = moralIntegration ~ log(heterogeneity), data = angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2419 -2.0135 -0.2082  2.4858  6.9491
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         22.7138     2.5478   8.915 3.82e-11 ***
## log(heterogeneity)  -3.5246     0.7678  -4.591 4.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.938 on 41 degrees of freedom
## Multiple R-squared:  0.3395, Adjusted R-squared:  0.3234
## F-statistic: 21.07 on 1 and 41 DF,  p-value: 4.141e-05
```

```r
# So, by itself, log(heterogeneity) explains the second most amount of
# variation, behind region.  Seems to me like region NEEDS to be included
# here.

m4a <- lm(moralIntegration ~ log(heterogeneity) * region, data = angell)
Anova(m4a)
```

```
## Anova Table (Type II tests)
##
## Response: moralIntegration
##                            Sum Sq Df F value    Pr(>F)
## log(heterogeneity)         24.286  1  5.1622   0.02934 *
```

```
## region                    185.734  3 13.1599 6.544e-06 ***
## log(heterogeneity):region   3.520  3  0.2494   0.86121
## Residuals                 164.658 35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m2, m4a)
```

```
## Analysis of Variance Table
##
## Model 1: moralIntegration ~ log(heterogeneity) + region
## Model 2: moralIntegration ~ log(heterogeneity) * region
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     38 168.18
## 2     35 164.66  3    3.5199 0.2494 0.8612
```

```
# interaction between these two final regressors not significant though.
# So, final model for now is m2.

m2poly <- lm(moralIntegration ~ log(heterogeneity) + region + I(log(heterogeneity)^2),
    data = angell)
summary(m2poly)
```

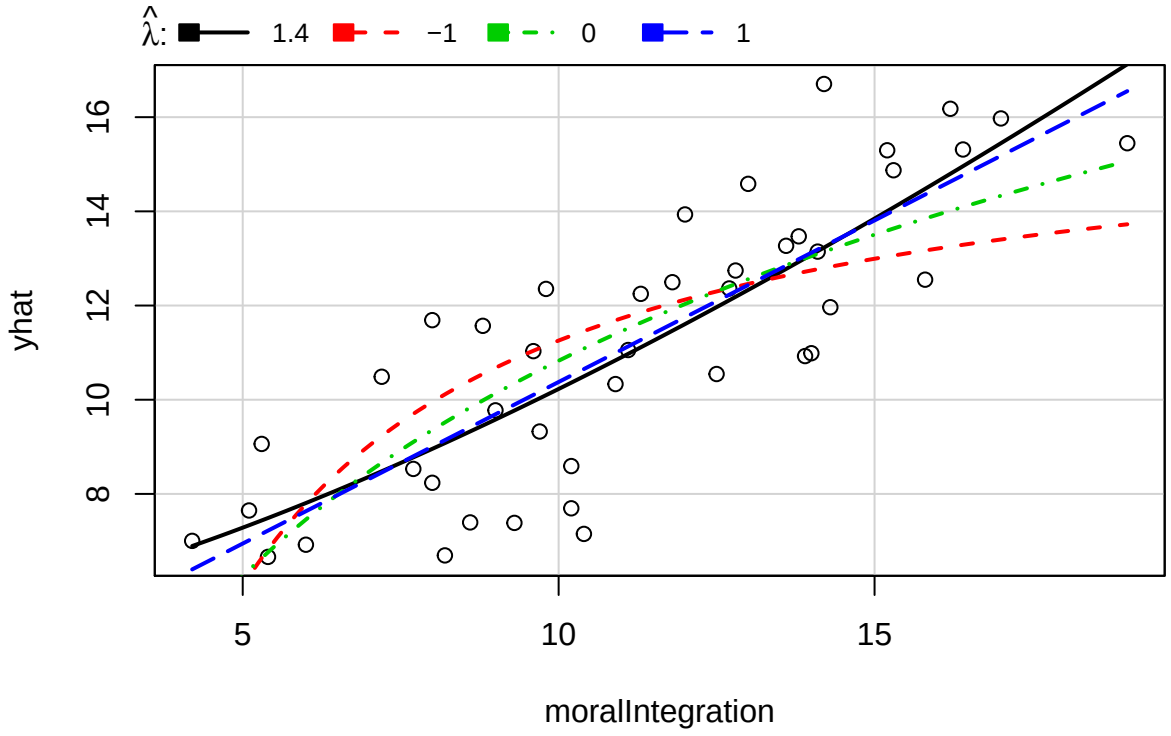
```
##
## Call:
## lm(formula = moralIntegration ~ log(heterogeneity) + region +
##     I(log(heterogeneity)^2), data = angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6619 -1.4628  0.0815  1.2770  3.5871
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               23.5496    11.1142   2.119  0.04088 *
## log(heterogeneity)        -3.3503     6.6743  -0.502  0.61867
## regionMW                  -3.2202     0.9121  -3.531  0.00113 **
## regionS                   -6.1626     1.1456  -5.379 4.35e-06 ***
## regionW                   -5.4134     1.1517  -4.700 3.55e-05 ***
## I(log(heterogeneity)^2)    0.2184     0.9925   0.220  0.82703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.131 on 37 degrees of freedom
## Multiple R-squared:  0.6865, Adjusted R-squared:  0.6442
## F-statistic: 16.21 on 5 and 37 DF,  p-value: 1.858e-08
```

```
# No need for the polynomial term


# Now, should we transform the predictor after we have this final model?

summary(powerTransform(m2))
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
```

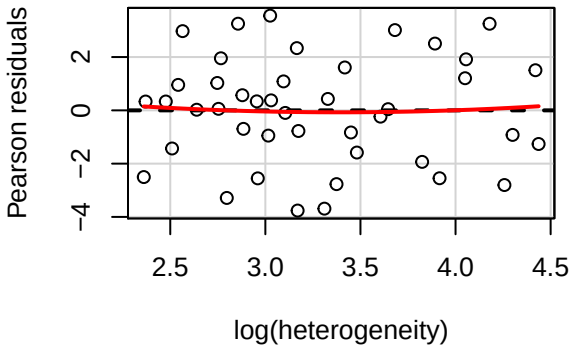
```
## Y1    1.1217           1       0.4361       1.8073
##
## Likelihood ratio tests about transformation parameters
##                               LRT df        pval
## LR test, lambda = (0) 10.7252197  1 0.001056851
## LR test, lambda = (1)  0.1220406  1 0.726831724
```

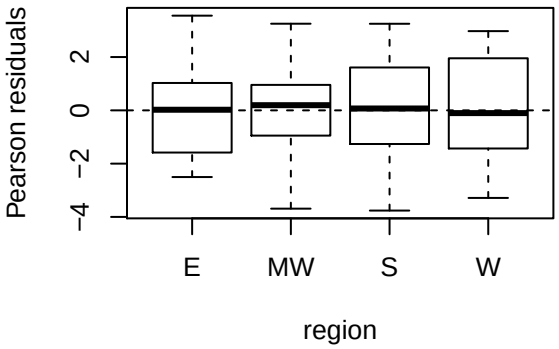
```
inverseResponsePlot(m2)
```

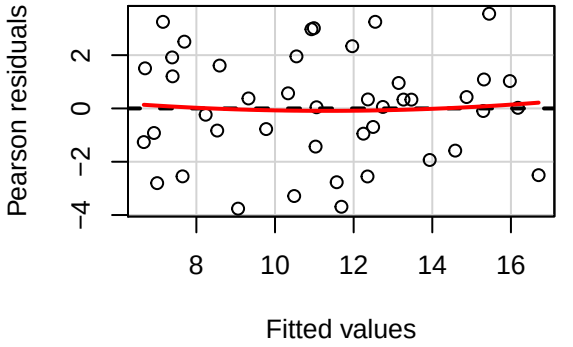



```
##      lambda      RSS
## 1  1.397786 113.7922
## 2 -1.000000 173.2182
## 3  0.000000 134.7576
## 4  1.000000 115.3920
```

```
# So, won't need to transform the response based on this.

# Should we include polynomials?

rp2 <- residualPlots(m2)
```








```
rp2
```

```
##                     Test stat Pr(>|t|)
## log(heterogeneity)      0.220    0.827
## region                     NA       NA
## Tukey test              0.495    0.621
```

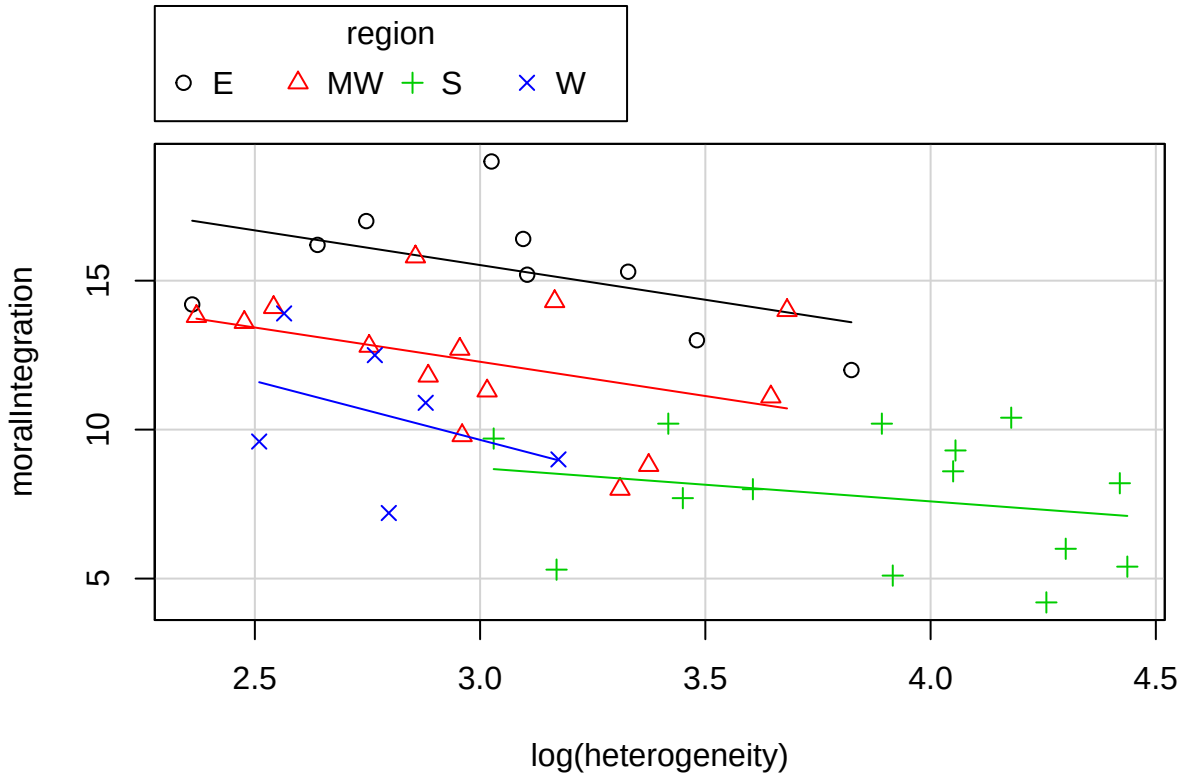
```
# So, no polynomial seems to be needed.

# Finally, let's look at the variance.

ncvTest(m2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.3184531    Df = 1     p = 0.5725388
```

```
# High p-value, can't reject null that the variance of this function is
# constant.
```

2.

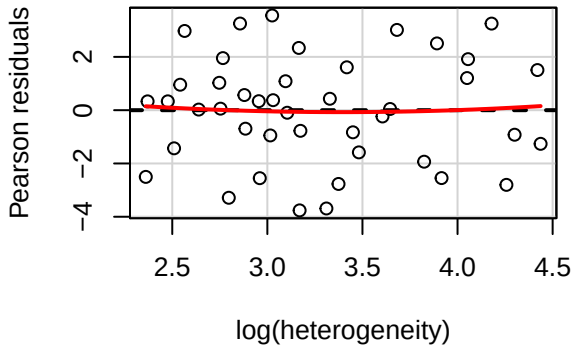
```
scatterplot(moralIntegration ~ log(heterogeneity) | region, data = angell, smooth = FALSE,
    boxplots = FALSE)
```

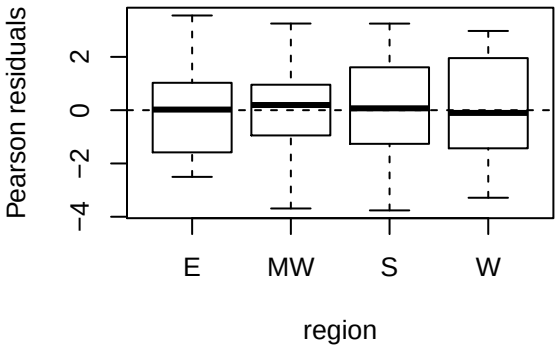


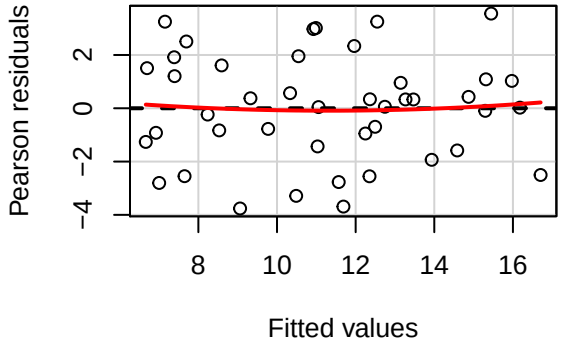

```
summary(m2)
```

```
##
## Call:
## lm(formula = moralIntegration ~ log(heterogeneity) + region,
##     data = angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7610 -1.3475  0.0564  1.3553  3.5536
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         21.1721     2.5756   8.220 5.87e-10 ***
## log(heterogeneity)  -1.8926     0.8079  -2.343 0.024493 *
## regionMW            -3.2169     0.9005  -3.572 0.000981 ***
## regionS             -6.1121     1.1083  -5.515 2.63e-06 ***
## regionW             -5.3904     1.1325  -4.760 2.80e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.104 on 38 degrees of freedom
## Multiple R-squared:  0.6861, Adjusted R-squared:  0.6531
## F-statistic: 20.77 on 4 and 38 DF,  p-value: 3.851e-09
```

```
residualPlots(m2)
```





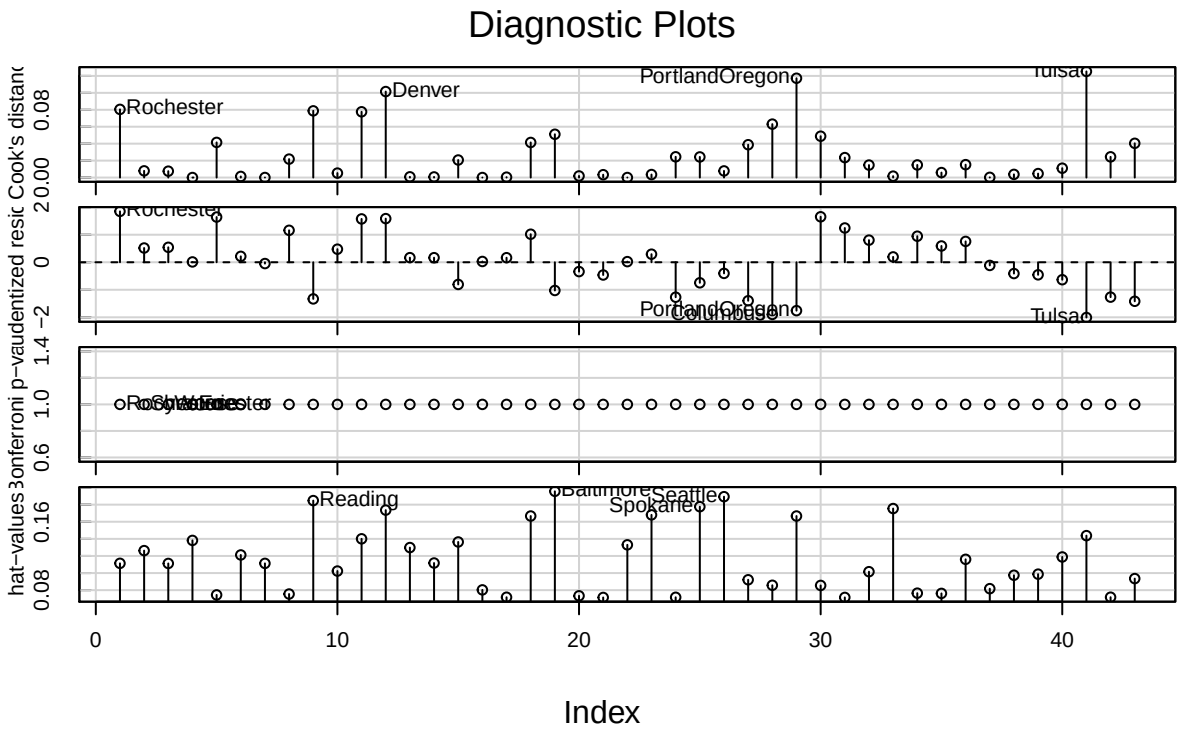



```
##                     Test stat Pr(>|t|)
## log(heterogeneity)      0.220    0.827
## region                     NA       NA
## Tukey test              0.495    0.621
```

```
# Again, the residuals look fine. Look like null-plots with no curvature
# and/or fan-shape.
```

3.

```
influenceIndexPlot(m2, id.n = 4)
```




```
# Directly testing for outliers
outlierTest(m2)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##        rstudent unadjusted p-value Bonferonni p
## Tulsa -2.007563           0.052035           NA
```

```
m2out <- update(m2, subset = -c(41, 29, 12, 1))

summary(m2)
```

```
##
## Call:
## lm(formula = moralIntegration ~ log(heterogeneity) + region,
##     data = angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7610 -1.3475  0.0564  1.3553  3.5536
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         21.1721     2.5756   8.220 5.87e-10 ***
## log(heterogeneity)  -1.8926     0.8079  -2.343 0.024493 *
## regionMW            -3.2169     0.9005  -3.572 0.000981 ***
```

```
## regionS             -6.1121     1.1083  -5.515 2.63e-06 ***
## regionW             -5.3904     1.1325  -4.760 2.80e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.104 on 38 degrees of freedom
## Multiple R-squared:  0.6861, Adjusted R-squared:  0.6531
## F-statistic: 20.77 on 4 and 38 DF,  p-value: 3.851e-09
```

```
summary(m2out)
```

```
##
## Call:
## lm(formula = moralIntegration ~ log(heterogeneity) + region,
##     data = angell, subset = -c(41, 29, 12, 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5897 -1.1695  0.1641  1.2619  3.2311
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          21.714      2.397   9.059 1.38e-10 ***
## log(heterogeneity)   -2.214      0.750  -2.952 0.005694 **
## regionMW             -2.796      0.829  -3.373 0.001868 **
## regionS              -5.105      1.054  -4.844 2.73e-05 ***
## regionW              -4.945      1.157  -4.274 0.000147 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.866 on 34 degrees of freedom
## Multiple R-squared:  0.7159, Adjusted R-squared:  0.6825
## F-statistic: 21.42 on 4 and 34 DF,  p-value: 6.737e-09
```