STAT-S631

Final Exam

John Koo

Statement

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: *John Koo, 12/11/2017*

Question 1

When performing transformations, I will try to limit the extent to which I transform the variables in order to make a more interpretable model. This may come at the cost of model performance.

Based on the scatterplot matrix (Fig 1), we can see moderate relationships between `moralIntegration` and all of the others without any particularly strong linear relationship between the two continuous predictors. There seems to be some relationship between `mobility` and the factor variable `region`, so just one or the other may be sufficient. Based on the scatterplots in the bottom row (with `moralIntegration` on the y-axis), there doesn't appear to be any reason to believe that a transformation is necessary. We will test this hypothesis.

First, we try power transformations on the continuous predictors with and without the factor variable. In either case, $\lambda = 0$ seems appropriate for `heterogeneity`, but the results disagree regarding `mobility` (see Tables 1, 2). Upon closer inspection, we can see that they both contain $\lambda = 0$. Testing $\lambda = 0$ and $\lambda = 1$ for the transformations without `region` shows that we achieve distributions closer to multivariate normal with $\lambda = 0$ for `mobility` (see Tables 3 and 4). So our final transformations will be log transformations for both continuous predictors.

Building a full model on these predictors, we can see that the pairwise interaction terms are not significant (Table 5). However, we cannot say that they are all insignificant from just this result. For that, we need a type I test. Table 6 shows that we cannot reject the null hypothesis that the interaction terms have zero coefficients at typical levels of significance (e.g., $\alpha < 0.1$). However, this is at a cost of model performance (R^2 decreases from ~0.8 to ~0.7).

Now that the interaction terms have been deemed insignificant, we can focus on the individual terms. A type II test on the parallel model shows that `log(mobility)` or `region` (or both) may not be significant (Table 7). This corresponds to our original observation from the scatterplots. Fitting a model without `region` and then another model without `log(mobility)` shows us that the model without `region` performs better in terms of R^2 and MSE. Then the final model is:

$$E[Y|X_1, X_2] = \beta_0 + \beta_1 \log(X_1) + \beta_2 \log(X_2)$$

Where Y is `moralIntegration` and the indices 1 and 2 refer to `heterogeneity` and `mobility` respectively. As a final check, we can compare this reduced model to the full model (Table 8). We can see that there is no significant difference here at usual levels of α (e.g., < 0.1).

Next, we can consider transformations on the response variable. A plot of log-likelihood vs λ for Box-Cox transformations shows that no transformation is necessary (Fig 2).

Finally, we can test our assumptions on the error terms by looking at the residuals. From a scatterplot of $\hat{e}$ vs $\hat{y}$, we see no reason to suspect that the assumption of constant variance is violated (Fig 3). The tests for non-constant variance confirm this. A Shapiro-Wilk test also confirms that the residuals are normally distributed (also, see Q-Q plot, Fig 4). No weights are needed.

At this point, polynomials will not be considered, since it appears that the model behaves normally and we have no reason to believe that any assumptions are broken.

Question 2

- $\hat{\beta}_0$: In the (theoretical) case where `heterogeneity` and `mobility` are 1, the average `moralIntegration` is 42.204.
- $\hat{\beta}_1$: For a fixed value of `mobility`, one unit change in `heterogeneity` on average decreases `moralIntegration` by 3.783 times the value of `heterogeneity`. This is because $\frac{\partial}{\partial x} A \log x = \frac{A}{x}$.
- $\hat{\beta}_2$: The interpretation here is the same as for $\hat{\beta}_1$. For a unit increase in `mobility` and fixed `heterogeneity`, `moralIntegration` decreases on average by 5.730 times the value of `mobility`.

From Question (1), we saw that there is no reason to believe that residuals are not normally distributed, correlated with the regressors, or auto-correlated. We also don't have evidence of non-constant variance. Since we have no reason to believe that our assumptions are violated, we will not be making any changes to the model (see Fig 5 and Table 9).

Question 3

From Fig 6, we can see that no outliers were detected based on Studentized residuals, with Bonferroni p-values all near 1. Based on Cook's distance, the most influential rows correspond to San Diego, Rochester, Portland (OR), and Houston (see Table 10). Removing these changes the coefficient estimates to 40.22, -3.784, and -5.128. An F-test comparing the new model to the old estimates shows that there is no significant difference (p-value = 0.95).

Accompanying code, outputs, and visualizations

```r
# packages, etc.
import::from(magrittr, `%>%`, `%<>%`)
dp <- loadNamespace('dplyr')
import::from(ggplot2, ggplot,
             geom_point,
             aes,
             theme_set, theme_bw,
             scale_colour_brewer, scale_x_log10,
             stat_smooth,
             labs)
import::from(GGally, ggpairs)
import::from(car, bcPower, boxCox,
             invTranPlot, invTranEstimate, invResPlot, powerTransform,
             ncvTest, residualPlots, influenceIndexPlot)
import::from(xtable, xtable)
import::from(car, Anova)
import::from(gridExtra, grid.arrange)
import::from(ggrepel, geom_label_repel)

theme_set(theme_bw())
```

```r
# load data
angell.df <- read.table('~/dev/stats-hw/stat-s631/Angell.txt') %>%
  dp$mutate(city = rownames(.))

summary(angell.df)
```

```
 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
     city
 Length:43
 Class :character
 Mode  :character
```

```r
# scatterplot matrix
angell.df %>%
  ggpairs(columns = c('region', 'heterogeneity', 'mobility',
                      'moralIntegration'),
```

```
         aes(colour = region))
```

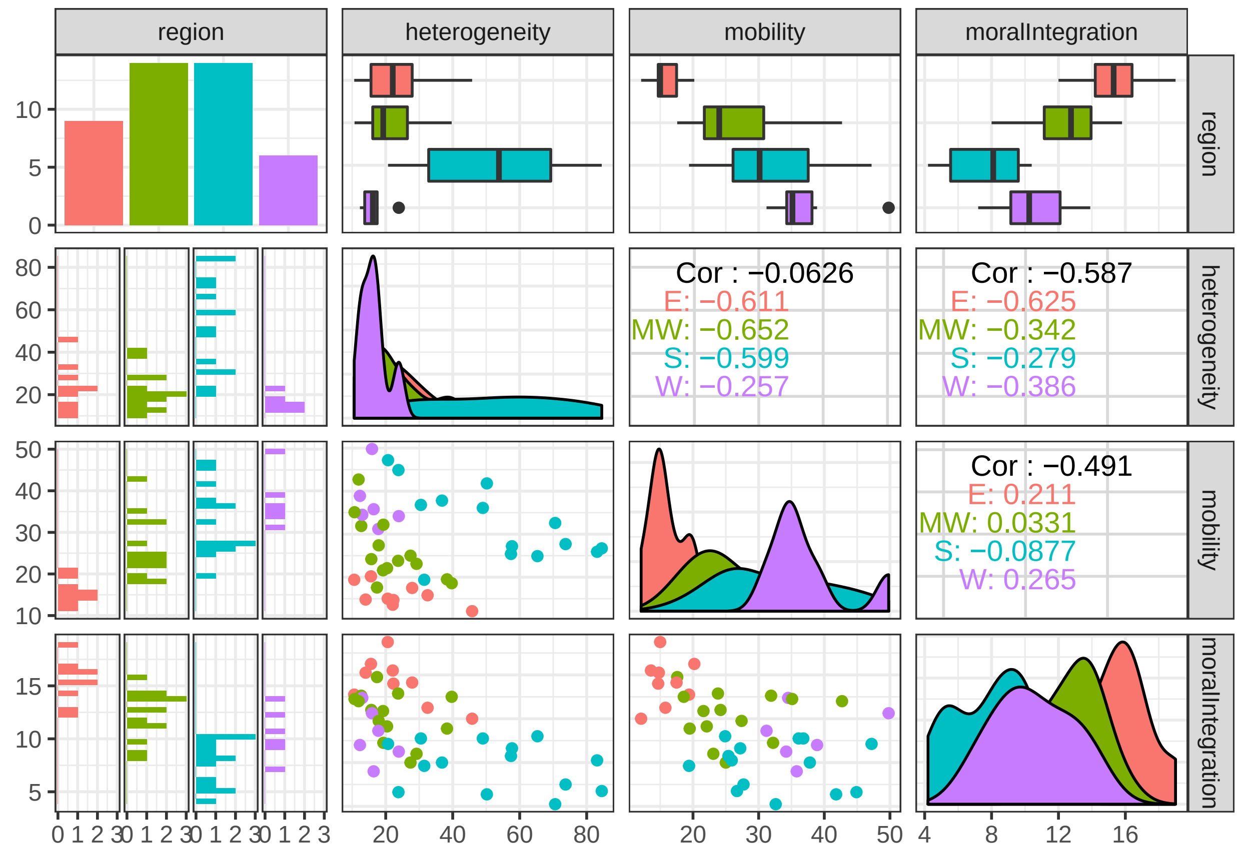

Figure 1: Scatterplot matrix

```
powerTransform(cbind(heterogeneity, mobility) ~ 1, angell.df) %>%
  summary() %>%
  .$result %>%
  xtable(caption = 'Transformations without the factor variable',
         label = 'tab:trans_wo_region') %>%
  print()
```

	Est Power	Rounded Pwr	Wald Lwr bnd	Wald Upr Bnd
heterogeneity	-0.42	0.00	-0.97	0.12
mobility	0.28	1.00	-0.57	1.12

Table 1: Transformations without the factor variable

```
powerTransform(cbind(heterogeneity, mobility) ~ region, angell.df) %>%
  summary() %>%
  .$result %>%
  xtable(caption = 'Transformations with the factor variable',
         label = 'tab:trans_w_region') %>%
```

```
  print()
```

	Est Power	Rounded Pwr	Wald Lwr bnd	Wald Upr Bnd
heterogeneity	-0.23	0.00	-0.67	0.20
mobility	-0.40	0.00	-1.07	0.26

Table 2: Transformations with the factor variable

```
powerTransform(cbind(heterogeneity, mobility) ~ 1, angell.df) %>%
  car::testTransform(c(0, 0)) %>%
  xtable(caption = paste('Results for transformations without the predictor',
                         'where both powers are 0'),
         label = 'tab:test_0') %>%
  print()
```

	LRT	df	pval
LR test, lambda = (0 0)	2.76	2	0.25

Table 3: Results for transformations without the predictor where both powers are 0

```
powerTransform(cbind(heterogeneity, mobility) ~ 1, angell.df) %>%
  car::testTransform(c(0, 1)) %>%
  xtable(caption = paste('Results for transformations without the predictor',
                         'where the power for mobility is 1'),
         label = 'tab:test_1') %>%
  print()
```

	LRT	df	pval
LR test, lambda = (0 1)	5.14	2	0.08

Table 4: Results for transformations without the predictor where the power for mobility is 1

```
full.mod <- lm(moralIntegration ~ log(heterogeneity) * log(mobility) * region,
               data = angell.df)
Anova(full.mod) %>%
  xtable(caption = 'Type II ANOVA on the full model',
         label = 'tab:full_model') %>%
  print()
```

	Sum Sq	Df	F value	Pr(>F)
log(heterogeneity)	37.04	1	8.92	0.0059
log(mobility)	12.71	1	3.06	0.0915
region	14.37	3	1.15	0.3454
log(heterogeneity):log(mobility)	5.15	1	1.24	0.2751
log(heterogeneity):region	10.68	3	0.86	0.4750
log(mobility):region	10.92	3	0.88	0.4651
log(heterogeneity):log(mobility):region	27.62	3	2.22	0.1090
Residuals	112.06	27		

Table 5: Type II ANOVA on the full model

```
parallel.mod <- lm(moralIntegration ~ log(heterogeneity) + log(mobility) + region,
                   data = angell.df)

anova(parallel.mod, full.mod) %>%
  xtable(caption = 'Type I ANOVA comparing the full model and parallel model',
         label = 'tab:parallel_mod') %>%
  print()
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	37	155.90				
2	27	112.06	10	43.84	1.06	0.4269

Table 6: Type I ANOVA comparing the full model and parallel model

```
Anova(parallel.mod) %>%
  xtable(caption = 'Type II ANOVA for the no-interaction model',
         label = 'tab:parallel_mod_2') %>%
  print()
```

	Sum Sq	Df	F value	Pr(>F)
log(heterogeneity)	36.54	1	8.67	0.0056
log(mobility)	12.28	1	2.91	0.0962
region	13.95	3	1.10	0.3599
Residuals	155.90	37		

Table 7: Type II ANOVA for the no-interaction model

```
final.mod <- lm(moralIntegration ~ log(heterogeneity) + log(mobility),
                data = angell.df)

anova(final.mod, full.mod) %>%
  xtable(caption = 'Type I ANOVA comparing the reduced model to the full model',
         label = 'tab:final_mod') %>%
  print()
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	40	169.85				
2	27	112.06	13	57.79	1.07	0.4213

Table 8: Type I ANOVA comparing the reduced model to the full model

```
boxCox(final.mod)
```

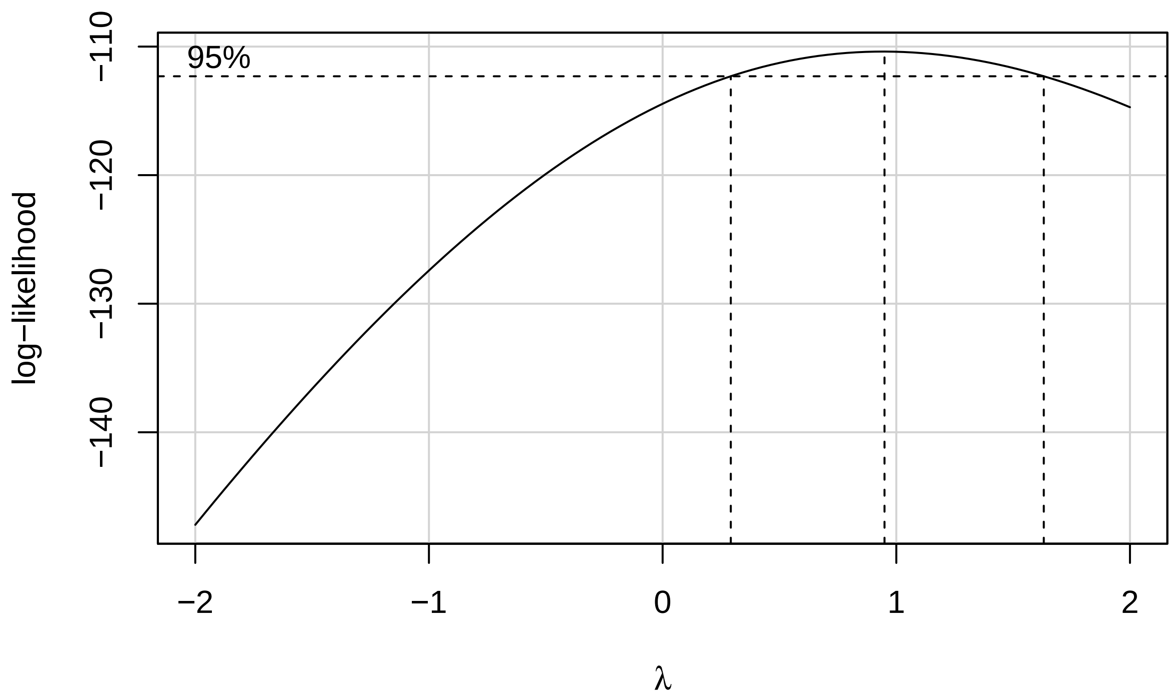

Figure 2: Log-likelihood for Box-Cox transformations

```
ggplot() +
  geom_point(aes(x = fitted.values(final.mod),
                 y = final.mod$residuals)) +
  labs(x = expression(hat(y)),
       y = expression(hat(e))) +
  stat_smooth(aes(x = fitted.values(final.mod),
                  y = final.mod$residuals),
              method = 'lm', formula = y ~ poly(x, 2, raw = TRUE))
```

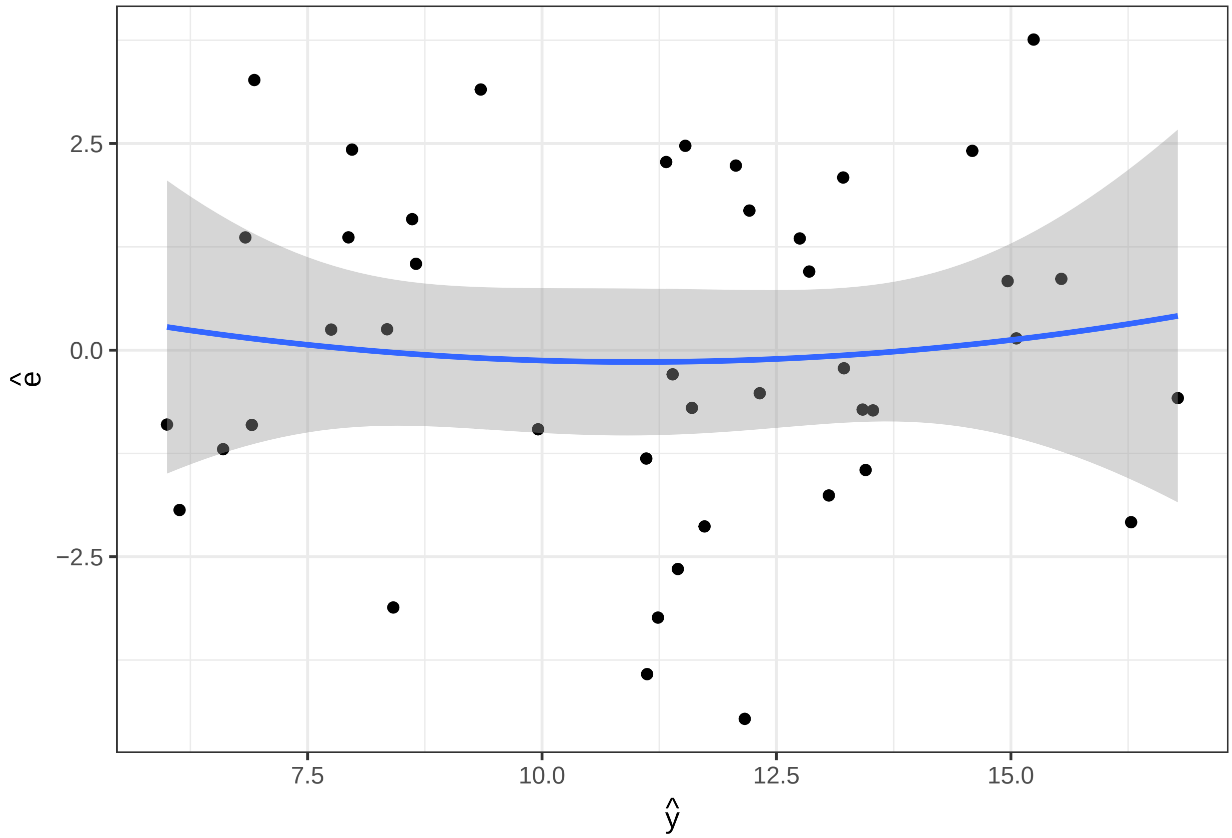

Figure 3: Residuals vs predicted values

```
ncvTest(final.mod)
```

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

```
ncvTest(final.mod, ~ region)
```

```
Non-constant Variance Score Test
Variance formula: ~ region
Chisquare = 1.011536    Df = 3     p = 0.7984607
```

```
ncvTest(final.mod, ~ heterogeneity)
```

```
Non-constant Variance Score Test
Variance formula: ~ heterogeneity
Chisquare = 0.330068    Df = 1     p = 0.565619
```

```
ncvTest(final.mod, ~ mobility)
```

```
Non-constant Variance Score Test
Variance formula: ~ mobility
Chisquare = 0.21742    Df = 1     p = 0.6410128
```

```r
ncvTest(final.mod, ~ log(heterogeneity))
```

```
Non-constant Variance Score Test
Variance formula: ~ log(heterogeneity)
Chisquare = 0.1046286    Df = 1     p = 0.7463443
```

```r
ncvTest(final.mod, ~ log(mobility))
```

```
Non-constant Variance Score Test
Variance formula: ~ log(mobility)
Chisquare = 0.1467762    Df = 1     p = 0.7016354
```

```r
qqnorm(final.mod$residuals)
qqline(final.mod$residuals)
```

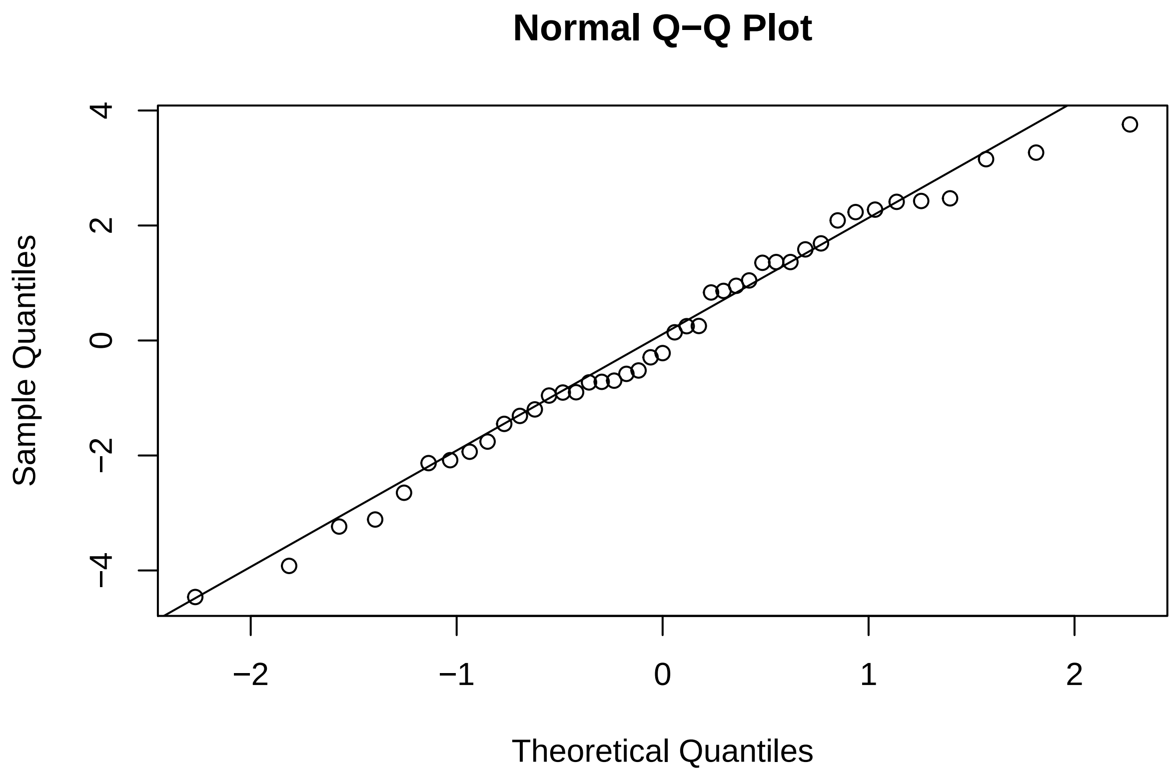

Figure 4: Q-Q plot of the residuals

```r
shapiro.test(final.mod$residuals)
```

```
    Shapiro-Wilk normality test

data:  final.mod$residuals
W = 0.98195, p-value = 0.7242
```

```r
summary(final.mod)
```

```

Call:
lm(formula = moralIntegration ~ log(heterogeneity) + log(mobility),
    data = angell.df)

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
```

```r
residualPlots(final.mod) %>%
  as.data.frame() %>%
  xtable(caption = 'Tukey tests',
         label = 'tab:tukey') %>%
  print()
```

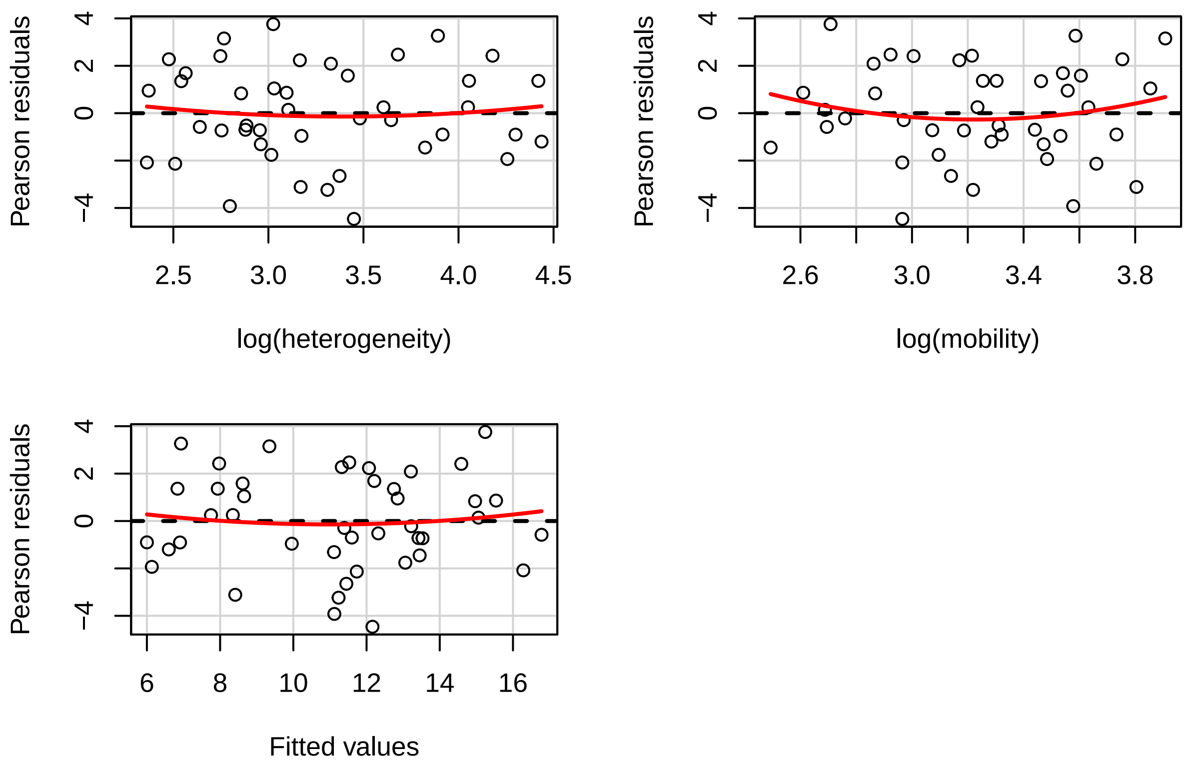

Figure 5: Residuals vs regressors

	Test stat	Pr(>|t|)
log(heterogeneity)	0.43	0.67
log(mobility)	0.90	0.37
Tukey test	0.49	0.62

Table 9: Tukey tests

```
influenceIndexPlot(final.mod)
```

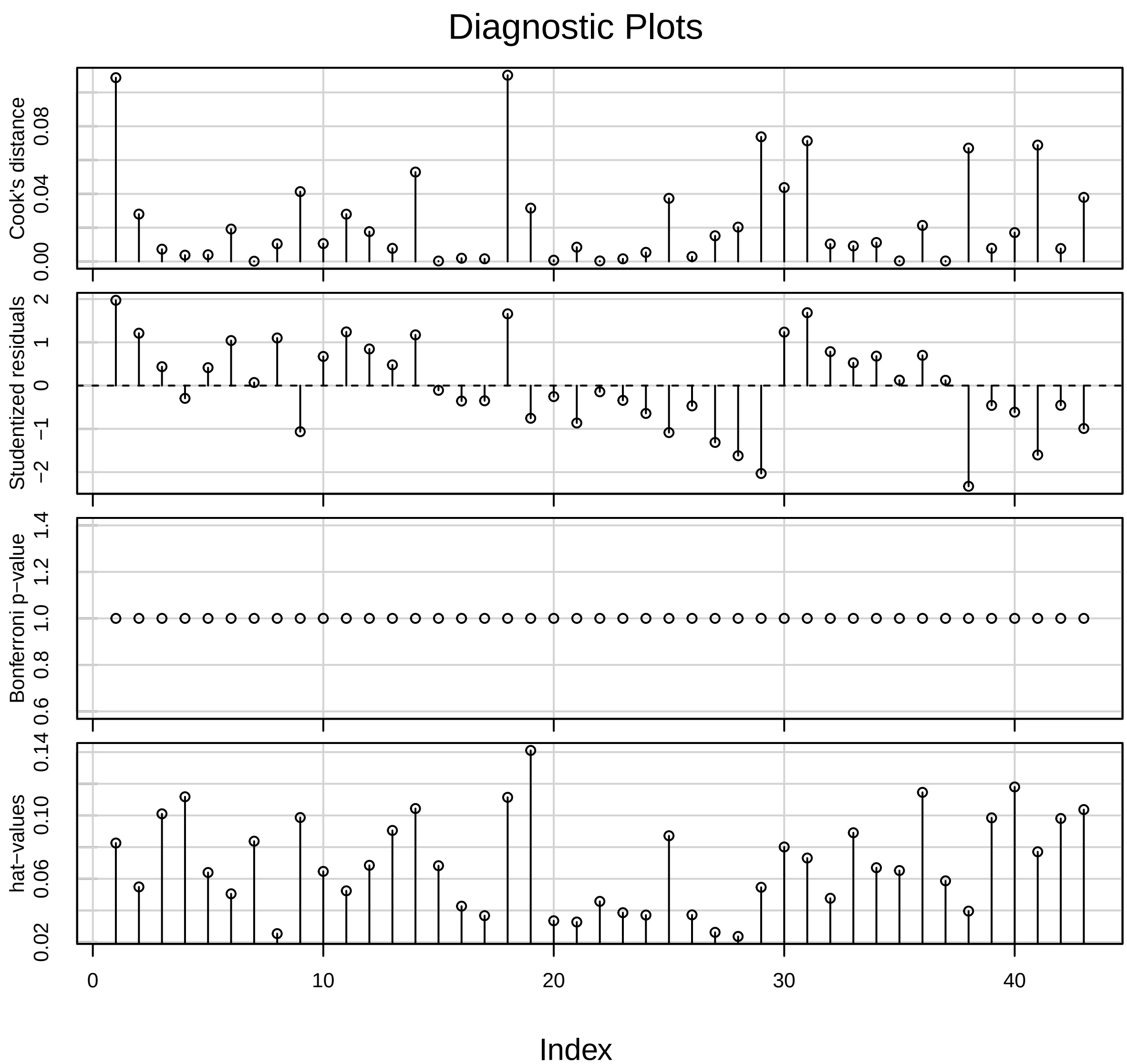

Figure 6: Influence index plot for the final model

```
angell.df %>%
  dp$mutate(cook = cooks.distance(final.mod)) %>%
  dp$arrange(-cook) %>%
  head() %>%
  dp$rename('Cook\'s distance' = cook) %>%
  xtable(caption = 'Most influential rows based on Cook\'s distance',
         label = 'tab:cook') %>%
  print(include.rownames = FALSE)
```

moralIntegration	heterogeneity	mobility	region	city	Cook's distance
12.50	15.90	49.80	W	SanDiego	0.11
19.00	20.60	15.00	E	Rochester	0.11
7.20	16.40	35.80	W	PortlandOregon	0.07
10.20	49.00	36.10	S	Houston	0.07
5.30	23.80	44.90	S	Tulsa	0.07
7.70	31.50	19.40	S	Louisville	0.07

Table 10: Most influential rows based on Cook's distance

```
angell.df %>%
  dp$mutate(resid = final.mod$residuals,
            yhat = fitted.values(final.mod)) %>%
  ggplot() +
  stat_smooth(aes(x = yhat, y = resid),
              method = 'lm', formula = y ~ poly(x, 2, raw = TRUE)) +
  geom_point(aes(x = yhat, y = resid)) +
  geom_label_repel(aes(x = yhat, y = resid, label = city)) +
  labs(x = expression(hat(y)),
       y = expression(hat(e)))
```

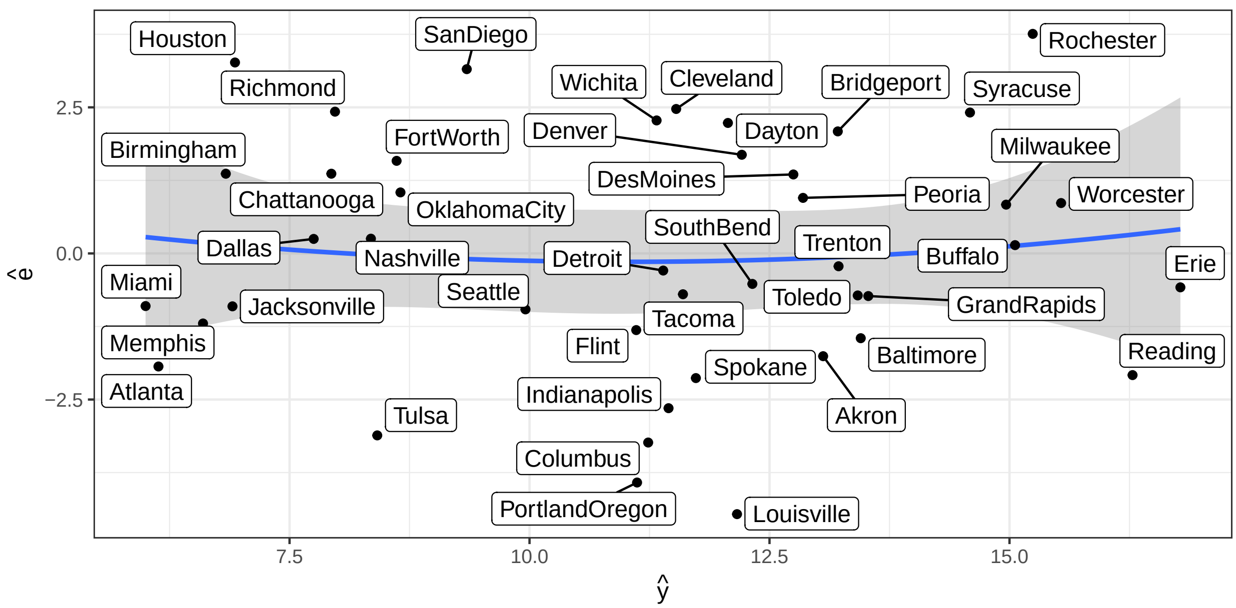

Figure 7: Residual plot with labels

```
influential.cities <- c('SanDiego', 'Rochester', 'PortlandOregon', 'Houston')

new.mod <- angell.df %>%
  dp$filter(!(city %in% influential.cities)) %>%
  lm(moralIntegration ~ log(heterogeneity) + log(mobility), data = .)
```

```r
summary(new.mod)
```

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

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
```

```r
# test for significant change
c. <- final.mod$coefficients
b <- new.mod$coefficients
V <- vcov(new.mod)
L <- diag(rep(1, 3))
q <- 3
F.stat <- t(L %*% b - c.) %*% solve(L %*% V %*% t(L)) %*% (L %*% b - c.) / q
1 - pf(F.stat, q, nrow(new.mod$model) - length(b))
```

```
          [,1]
[1,] 0.9512404
```