

Methods for the Association between
Segregation and Health in the
Northeastern and Southeastern
Brazilian metropolitan regions and
medium size cities

Antonio Ponce de Leon

Departamento de Epidemiologia
Instituto de Medicina Social
Universidade do Estado do Rio de Janeiro

Overview

- Introduction
- Material
- Model structure
- Modeling strategy
- Results
- Discussion

Introduction

Main hypothesis: Income segregation mediates the relationship between income, together with **income inequality**, and health outcomes, e.g. total mortality;

Data: Outcome is SMR (base = Brazil). Covariates are average income, income inequality indicators, and segregation indexes;

Level of analysis: Municipalities / Metropolitan Regions of the Southeast and Northeast.

Material

- Data on income come from Census 2000 **sample** \Rightarrow 10% of households, representative at the level of **conglomerate** (a set of census tracks);
- Households selected in the sample \Rightarrow **all income** received by everyone in the household;
- **Average** income at the conglomerate level.

Material

- Study units (**super units**) are municipalities or great metropolitan regions (MR), e.g. Rio de Janeiro 's MR consists of more than 10 municipalities, some **rather large**;
- Each **conglomerate** is a subunit within a superunit \Rightarrow income levels are representative of conglomerates;
- Levels of income are **averaged** or added up over conglomerates to the level of superunits.

Model structure

- Multiple linear regression with **weights** proportional to the observation (municipality or metropolitan region) population size;
- **Counts** of deaths are very large, so no need to Poisson or similar assumptions;
- Special care with **outliers**;
- Perhaps, great metropolitan regions should have been **split** into separate municipalities.

Model structure

Problem regarding **choice** of study units:

- (i) If units are **municipalities**, clusters would be formed for those within great metropolitan regions;
- (ii) If units consist of a **mix** of municipalities and metropolitan regions, one could think of a mix of two populations, from which samples of unequal sizes are extracted.

Modeling strategy

- Examine **visually** relationships between outcome and covariates;
- Problem \Rightarrow study units vary in population size (**bubble** plot and similar techniques);
- Examine normality of the data, using **Box-Cox** transformations;
- Theoretical model is clear, however there are **several** ways to regard it with variables at hand.

Modeling strategy

- 1st step (**all in**) \Rightarrow fit full model;
- 2nd step (**avoid redundancy**) \Rightarrow select variables out, according to the VIF (Variance Inflation Factor) up to a given tolerance (mean VIF <10);
- 3rd step (**keeping statistical significance**) \Rightarrow select variables out because they are not significant;
- 4th step (**avoid inconsistency**) \Rightarrow select variables out according to lack of plausibility.

Modeling strategy

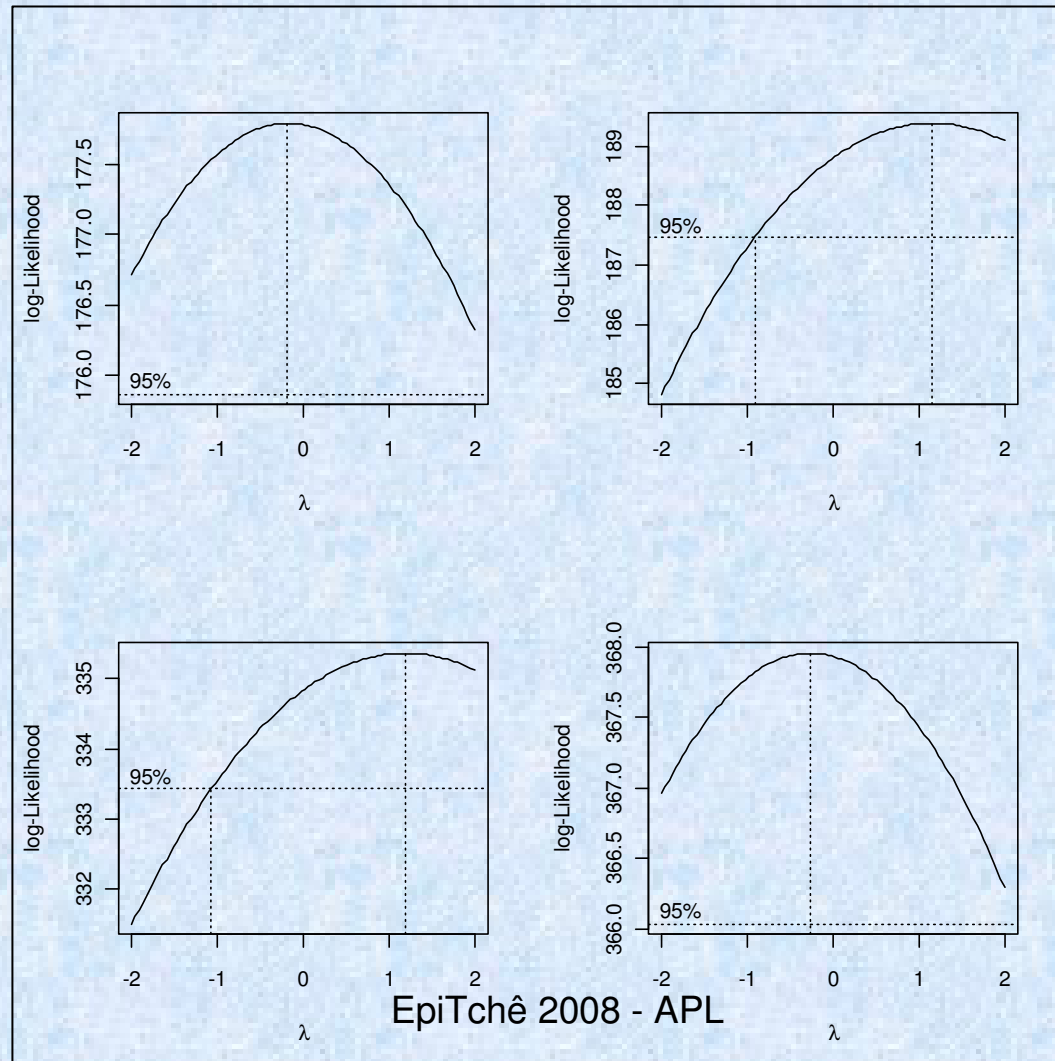
- Residual diagnostics focusing on **leverage** and influential values (outliers);
- After going for a model, reexamine normality of the data using **Box-Cox** transformations;
- Refit model if necessary;
- Different conclusions? Different significance levels?
- Magnitude and interpretation of effect are of importance?

Results

Four models were fitted:

- Men in Northeast \Rightarrow log transformation;
- Women in Northeast \Rightarrow no transformation;
- Men in Southeast \Rightarrow no transformation / inverse transformation;
- Women in Southeast \Rightarrow log transformation / no transformation.

Transformations for the dependent variable (univariate): Box-Cox statistics



Model for men in the Northeast

```
Call:  
lm(formula = smr_m2 ~ rendamd + perc_10r + r_40p, weights = pop_3a_m)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-1.60023	-0.39031	0.05993	0.39552	2.18855

```
Coefficients:
```

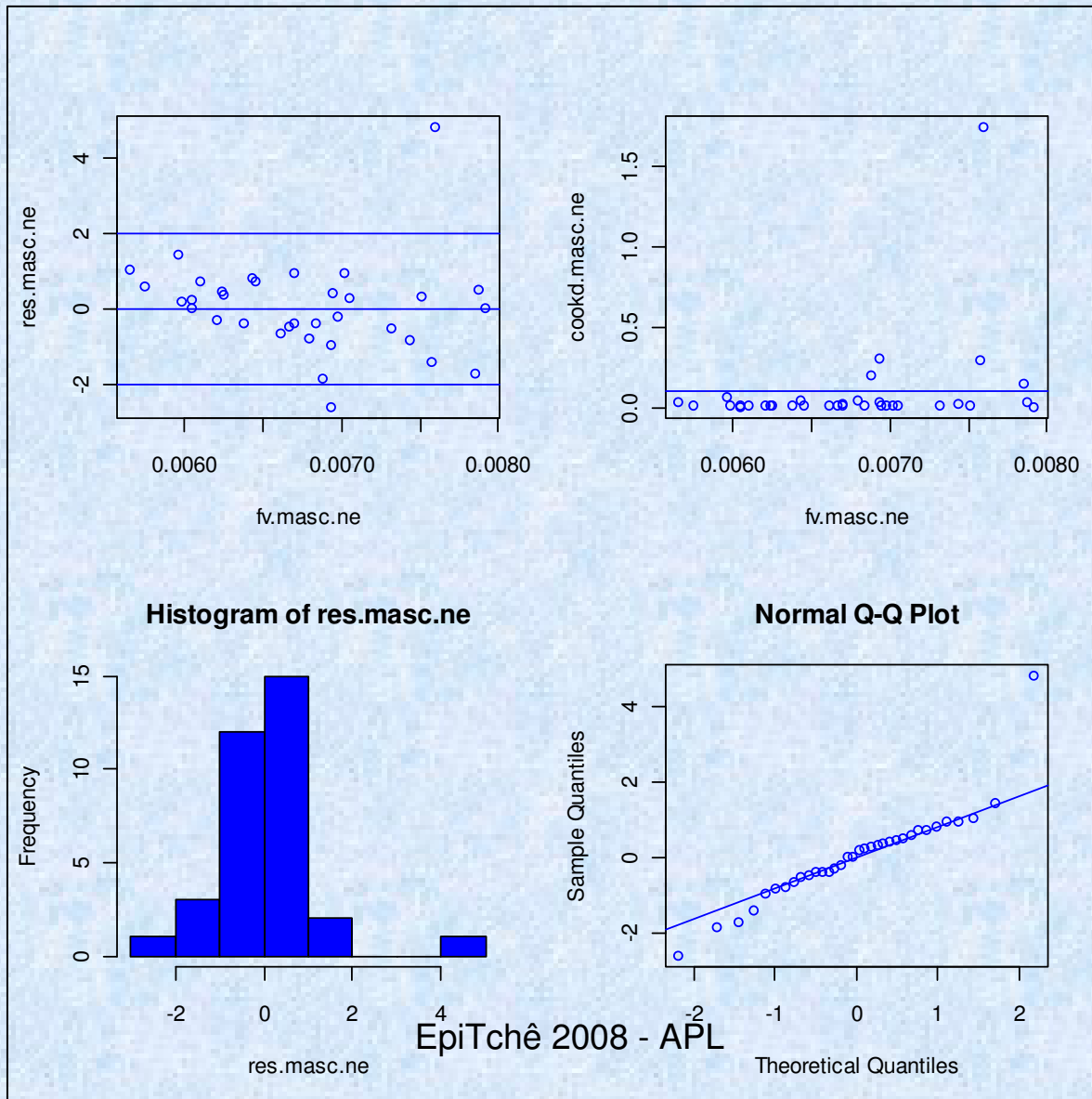
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.416e-02	3.731e-03	3.797	0.000665	***
rendamd	-1.710e-06	3.729e-06	-0.458	0.649899	
perc_10r	-2.852e-04	1.018e-04	-2.801	0.008822	**
r_40p	2.677e-04	8.434e-05	3.174	0.003463	**

```
---
```

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

```
Residual standard error: 0.7409 on 30 degrees of freedom  
Multiple R-squared: 0.342, Adjusted R-squared: 0.2761  
F-statistic: 5.196 on 3 and 30 DF, p-value: 0.005201
```

Residual diagnostics, all observations



Model for men in the Northeast, no outliers

```
Call:  
lm(formula = smr_m2 ~ rendamd + perc_10r + r_40p, weights = pop_3a_m)
```

Residuals:

Min	1Q	Median	3Q	Max
-197.42	-46.70	17.84	62.45	136.89

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-4.7597647	0.4511318	-10.551	1.93e-11	***
rendamd	-0.0001462	0.0004218	-0.347	0.7313	
perc_10r	-0.0158965	0.0125383	-1.268	0.2149	
r_40p	0.0207786	0.0101555	2.046	0.0499	*

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

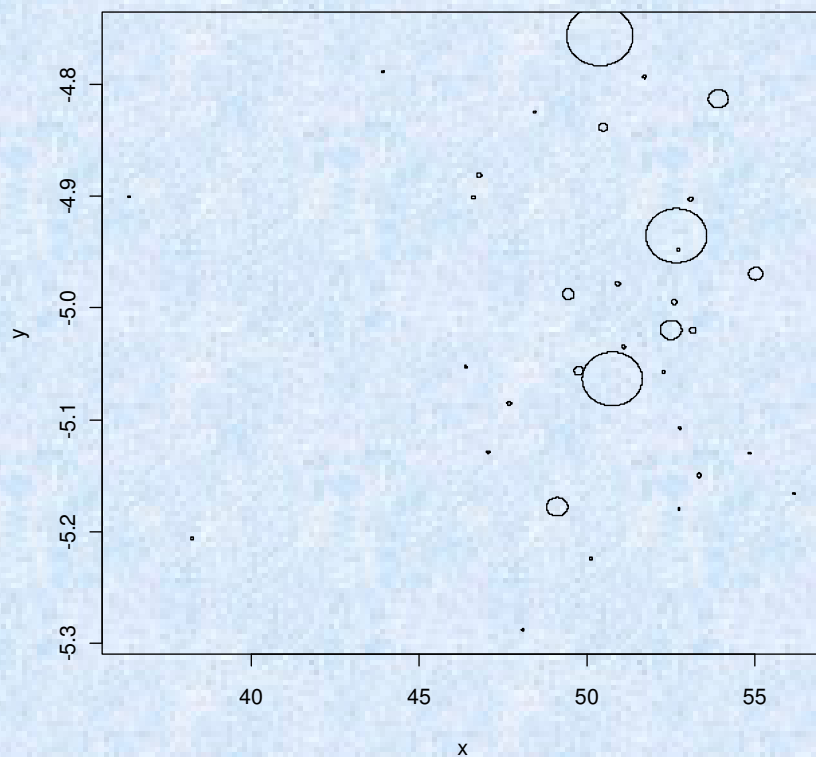
Residual standard error: 83.8 on 29 degrees of freedom

Multiple R-squared: 0.2083, Adjusted R-squared: 0.1264

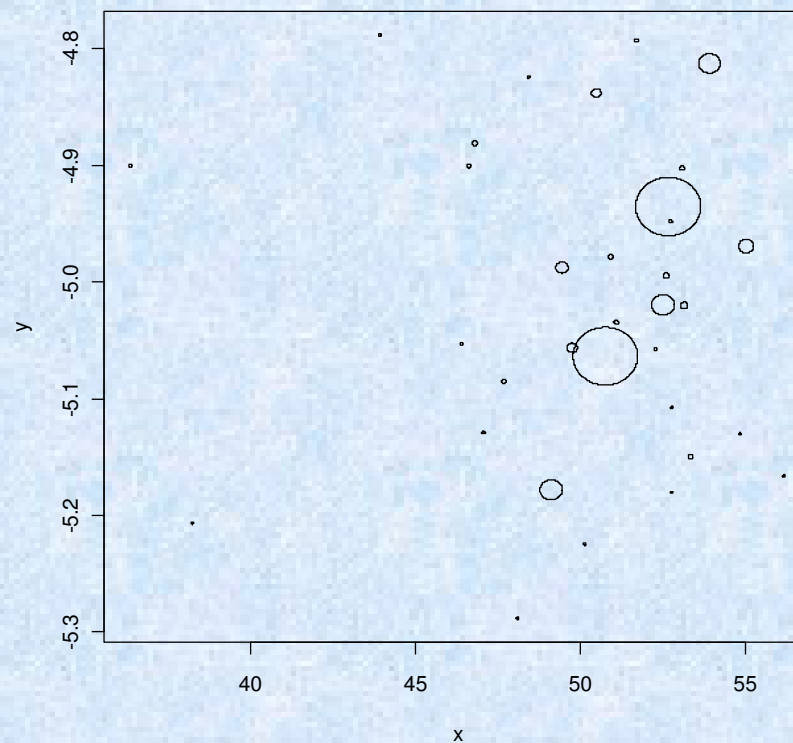
F-statistic: 2.543 on 3 and 29 DF, p-value: 0.07561

The difference it (Recife MR) makes

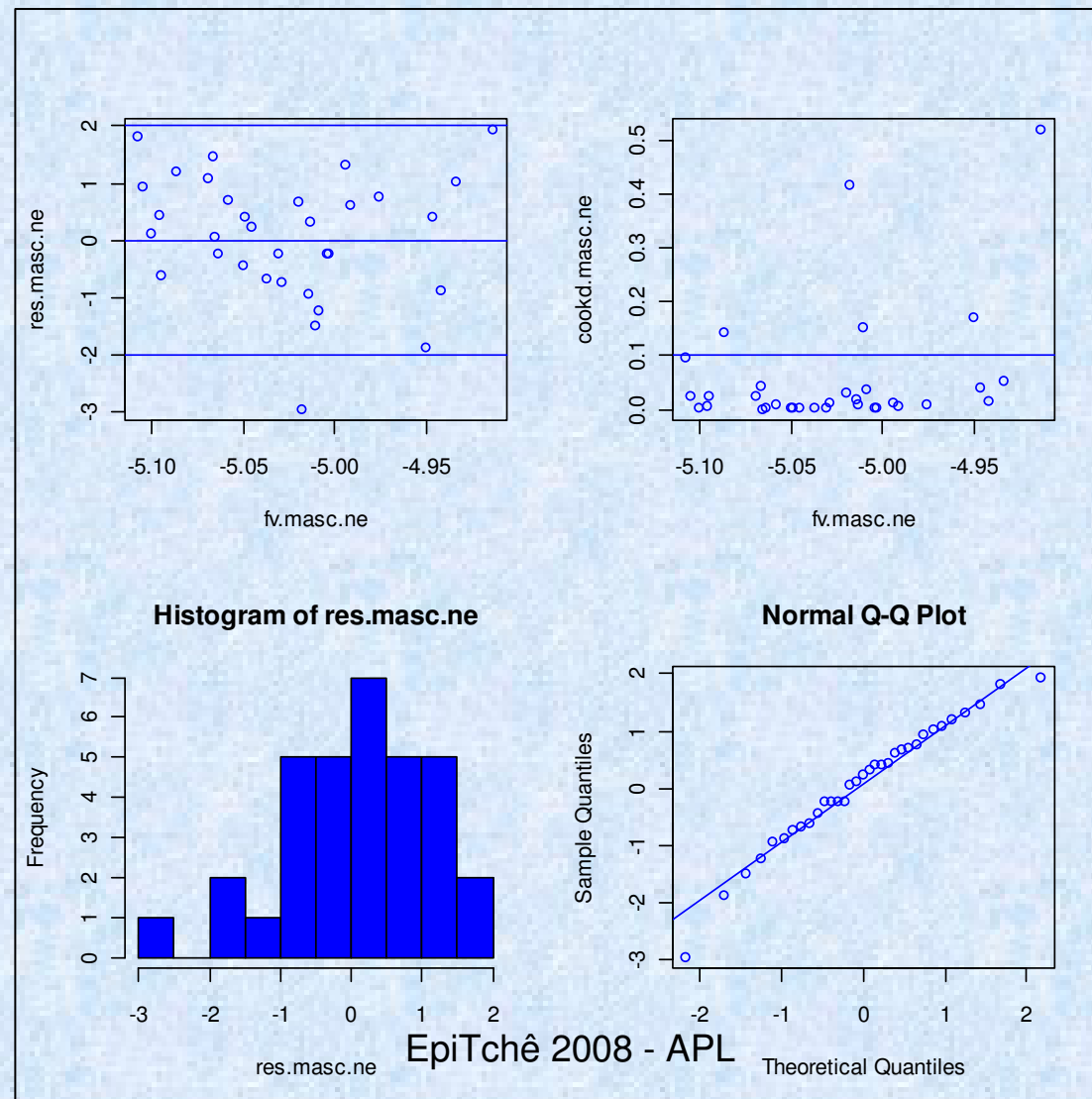
relationship between smr_m2 and perc_10r



relationship between smr_m2 and perc_10r



Residual diagnostics, no outliers (those for Cook's distance > 0.5)



Model for women in the Northeast

```
Call:
lm(formula = smr_f ~ rendamd + perc_10r + r_40p, weights = pop_3a_f)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.94383	-0.17108	0.05564	0.24466	0.74198

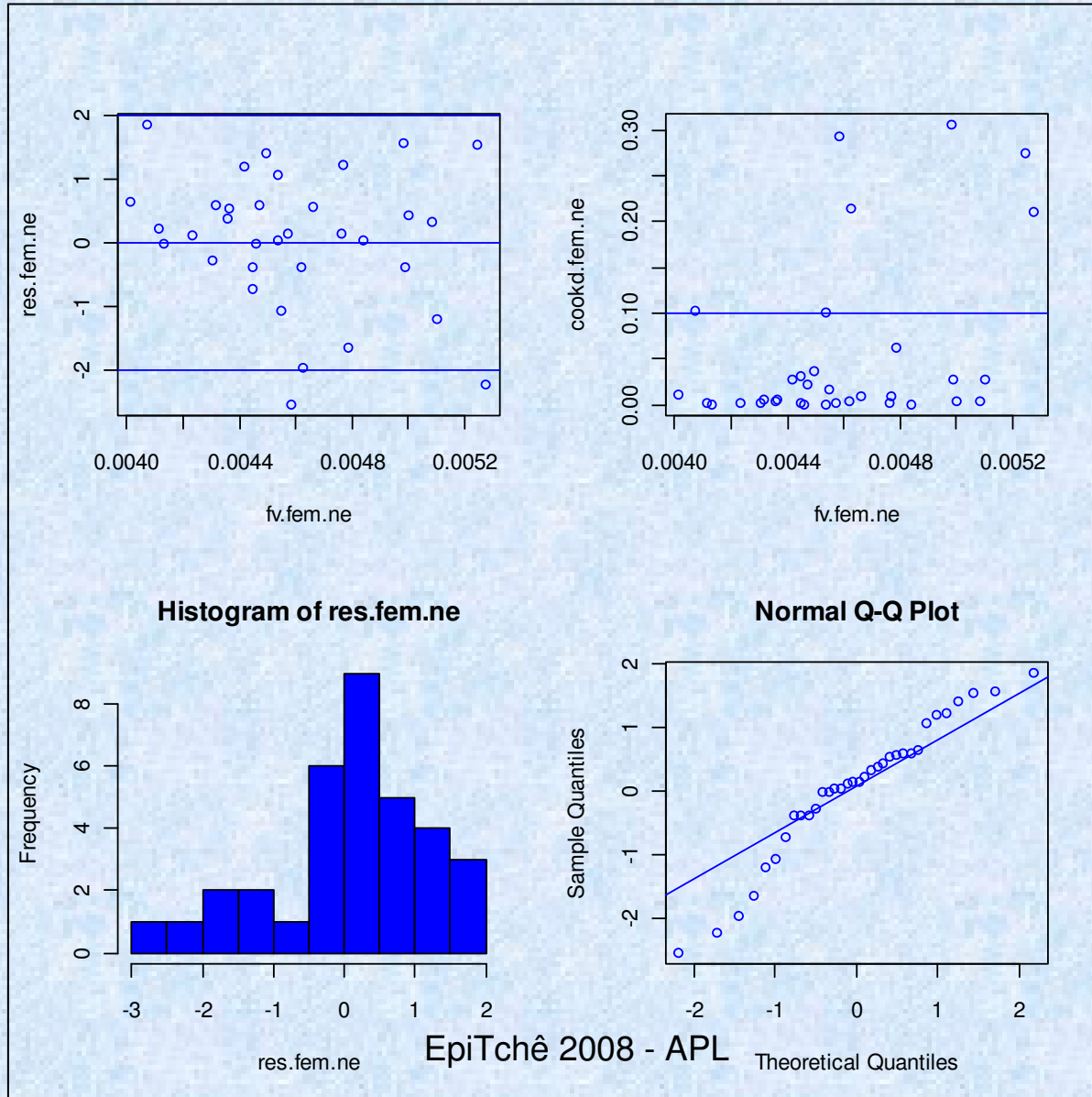
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	7.752e-03	2.147e-03	3.611	0.00110	**
rendamd	-1.948e-06	2.135e-06	-0.913	0.36878	
perc_10r	-1.346e-04	5.841e-05	-2.304	0.02830	*
r_40p	1.473e-04	4.820e-05	3.055	0.00469	**

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

Residual standard error: 0.4426 on 30 degrees of freedom
Multiple R-squared: 0.2957, Adjusted R-squared: 0.2253
F-statistic: 4.199 on 3 and 30 DF, p-value: 0.01356

Residual diagnostics, all observations



Model for men in the Southeast

```
Call:
lm(formula = smr_m ~ rendamd + rper_1q + int_pobr + npr, weights =
pop_3a_m)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.80763	-0.16129	0.08237	0.24101	1.70174

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.657e-03	8.166e-04	6.927	3.59e-09	***
rendamd	2.414e-07	2.107e-06	0.115	0.90920	
rper_1q	-2.359e-05	8.874e-06	-2.658	0.01010	*
int_pobr	6.851e-05	2.462e-05	2.783	0.00722	**
npr	-2.813e-03	1.236e-03	-2.275	0.02653	*

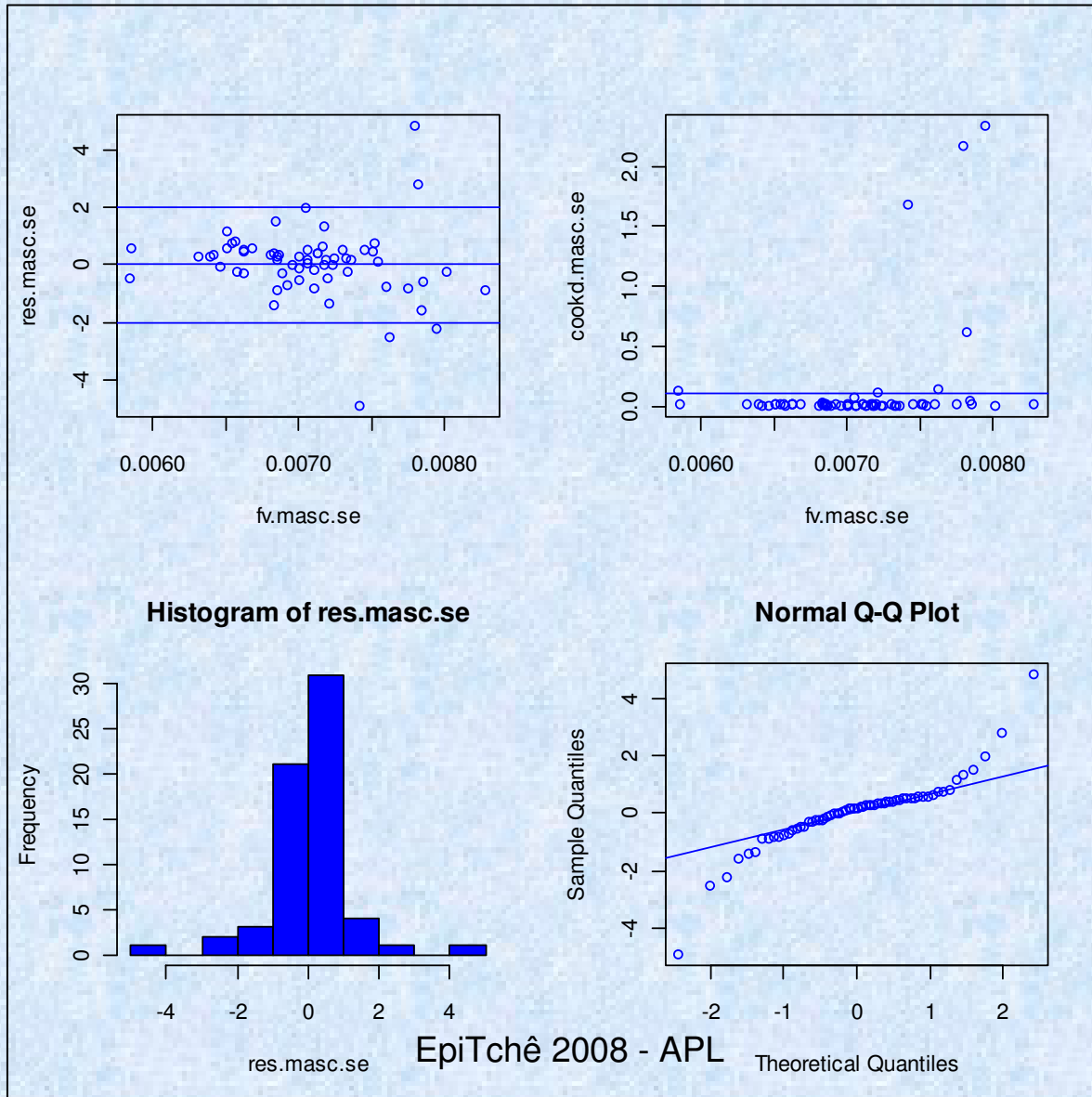
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5272 on 59 degrees of freedom

Multiple R-squared: 0.4571, Adjusted R-squared: 0.4203

F-statistic: 12.42 on 4 and 59 DF, p-value: 2.167e-07

Residual diagnostics, all observations



Model for men in the Southeast, no outliers

```
Call:
lm(formula = smr_m ~ rendamd + rper_1q + int_pobr + npr, weights =
pop_3a_m)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.92859	-0.16819	0.04512	0.18483	0.83712

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	8.622e-03	8.842e-04	9.751	1.37e-13	***
rendamd	4.871e-06	2.207e-06	2.207	0.031471	*
rper_1q	-3.885e-05	1.083e-05	-3.586	0.000713	***
int_pobr	-2.598e-05	2.488e-05	-1.044	0.300905	
npr	-1.574e-03	8.375e-04	-1.880	0.065427	.

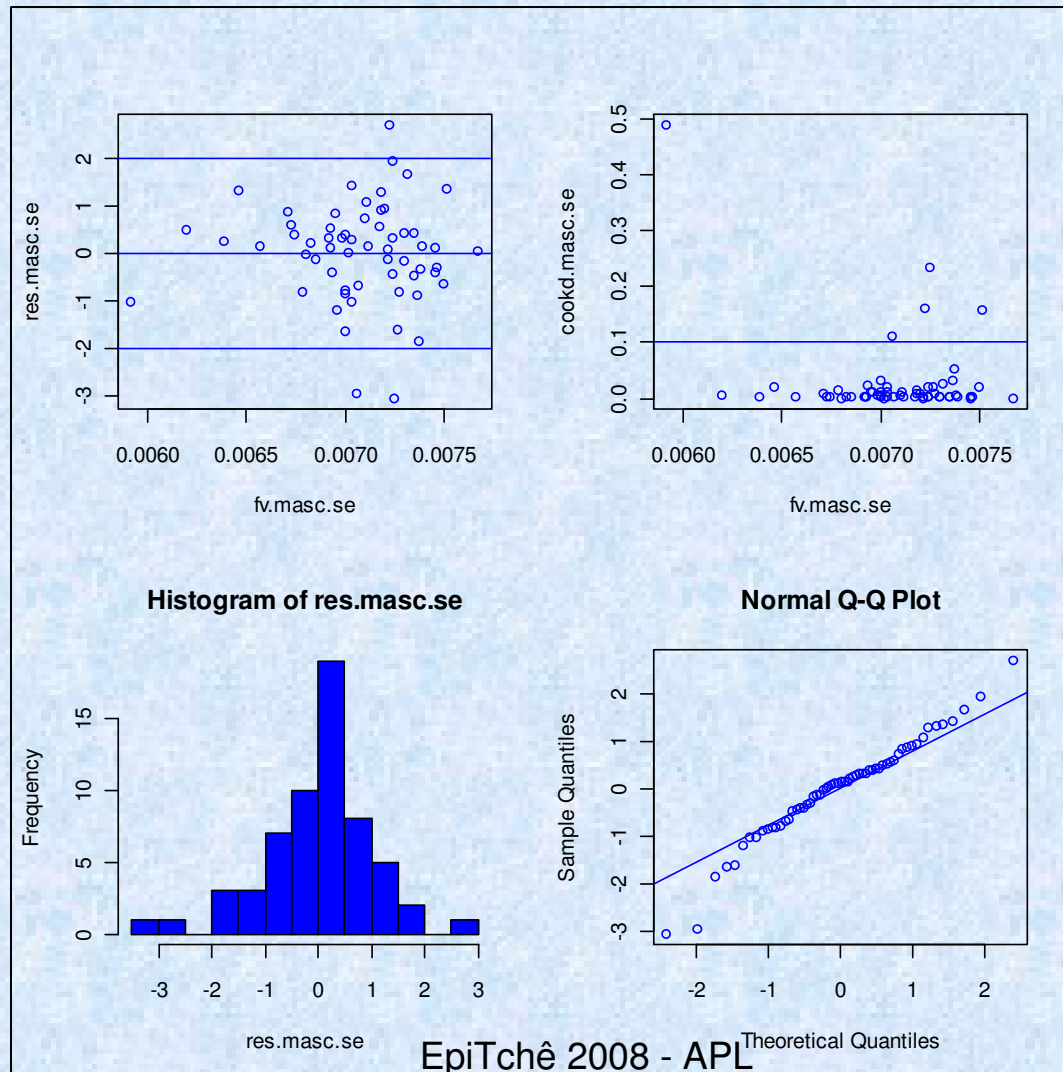
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3468 on 55 degrees of freedom

Multiple R-squared: 0.2373, Adjusted R-squared: 0.1818

F-statistic: 4.278 on 4 and 55 DF, p-value: 0.004379

Residual diagnostics, no outliers (those for Cook's distance > 0.5)



Model for women in the Southeast

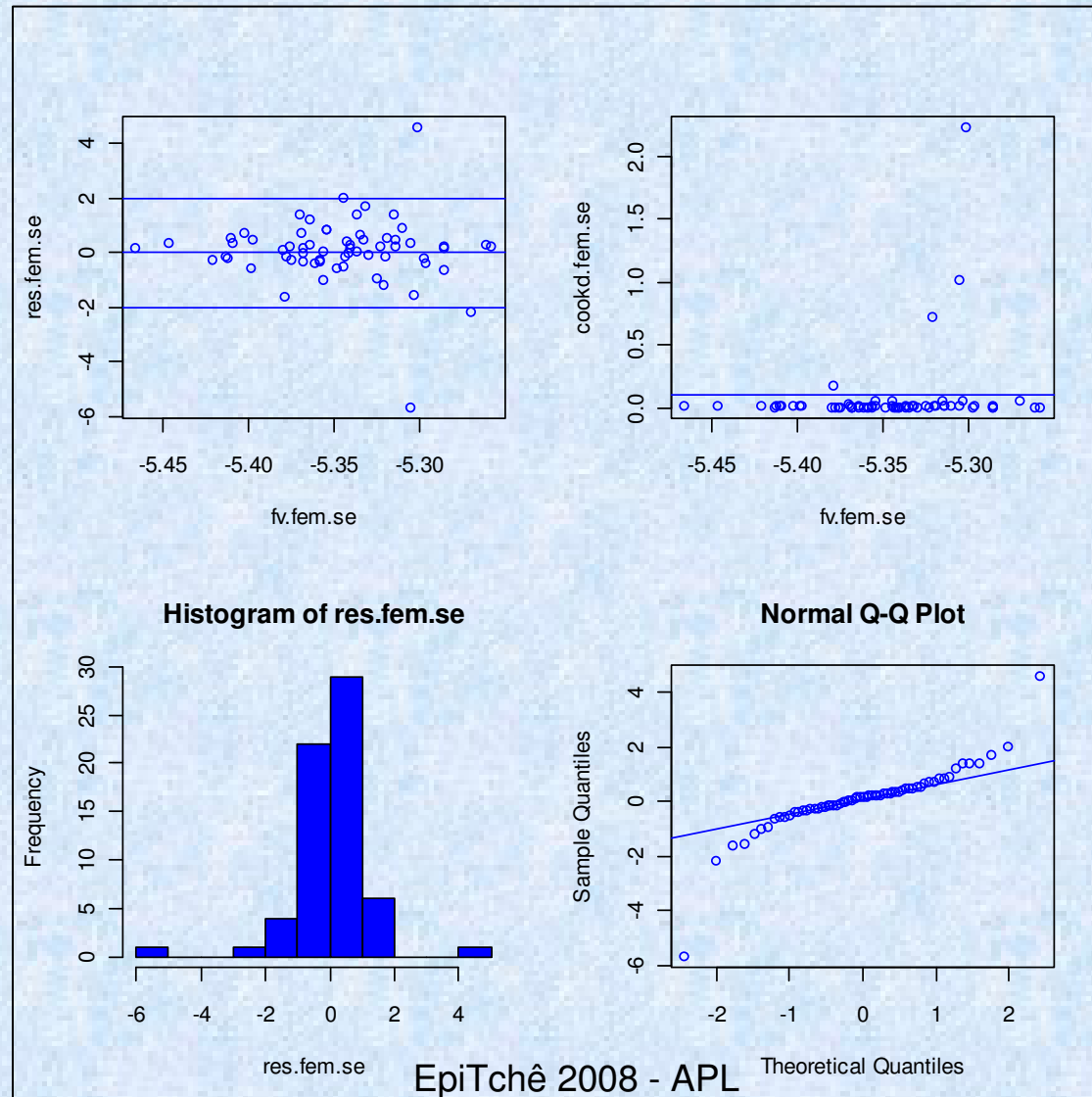
```
Call:
lm(formula = smr_f2 ~ rendamd + rper_lq + npr, weights = pop_3a_f)

Residuals:
      Min       1Q   Median       3Q      Max
-266.848  -17.739    9.057   28.852  201.115

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.174e+00  6.168e-02 -83.890  < 2e-16 ***
rendamd      8.654e-05  1.304e-04   0.664  0.509316
rper_lq     -3.398e-03  8.420e-04  -4.035  0.000157 ***
npr         -2.698e-01  1.395e-01  -1.934  0.057859 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 63.13 on 60 degrees of freedom
Multiple R-squared:  0.2259,    Adjusted R-squared:  0.1872
F-statistic: 5.835 on 3 and 60 DF,  p-value: 0.001446
```

Residual diagnostics, all observations



Model for women in the Southeast, no outliers

```
Call:
lm(formula = smr_f2 ~ rendamd + rper_1q + npr, weights = pop_3a_f)
```

Residuals:

Min	1Q	Median	3Q	Max
-135.40	-21.76	6.81	23.79	123.67

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-5.1436997	0.0523873	-98.186	<2e-16	***
rendamd	-0.0001530	0.0001998	-0.766	0.4469	
rper_1q	-0.0022521	0.0011437	-1.969	0.0538	.
npr	-0.2698995	0.1064576	-2.535	0.0140	*

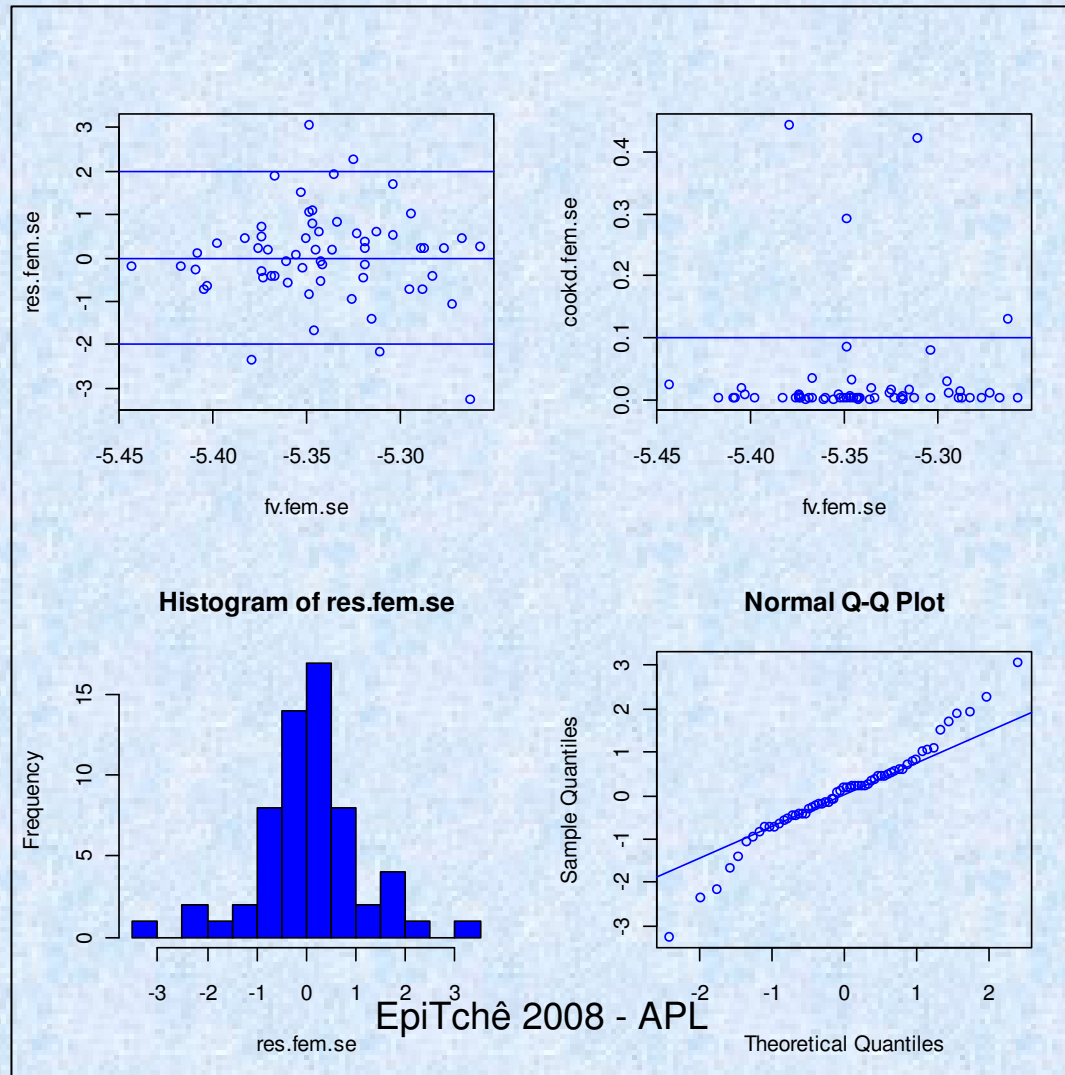
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.26 on 57 degrees of freedom

Multiple R-squared: 0.2362, Adjusted R-squared: 0.196

F-statistic: 5.877 on 3 and 57 DF, p-value: 0.001441

Residual diagnostics, no outliers (those for Cook's distance > 0.5)



Box-Cox transformations, based on given sets of independent variables

Northeast

```
boxcox(smr_m ~ rendamd + perc_10r + r_40p)
```

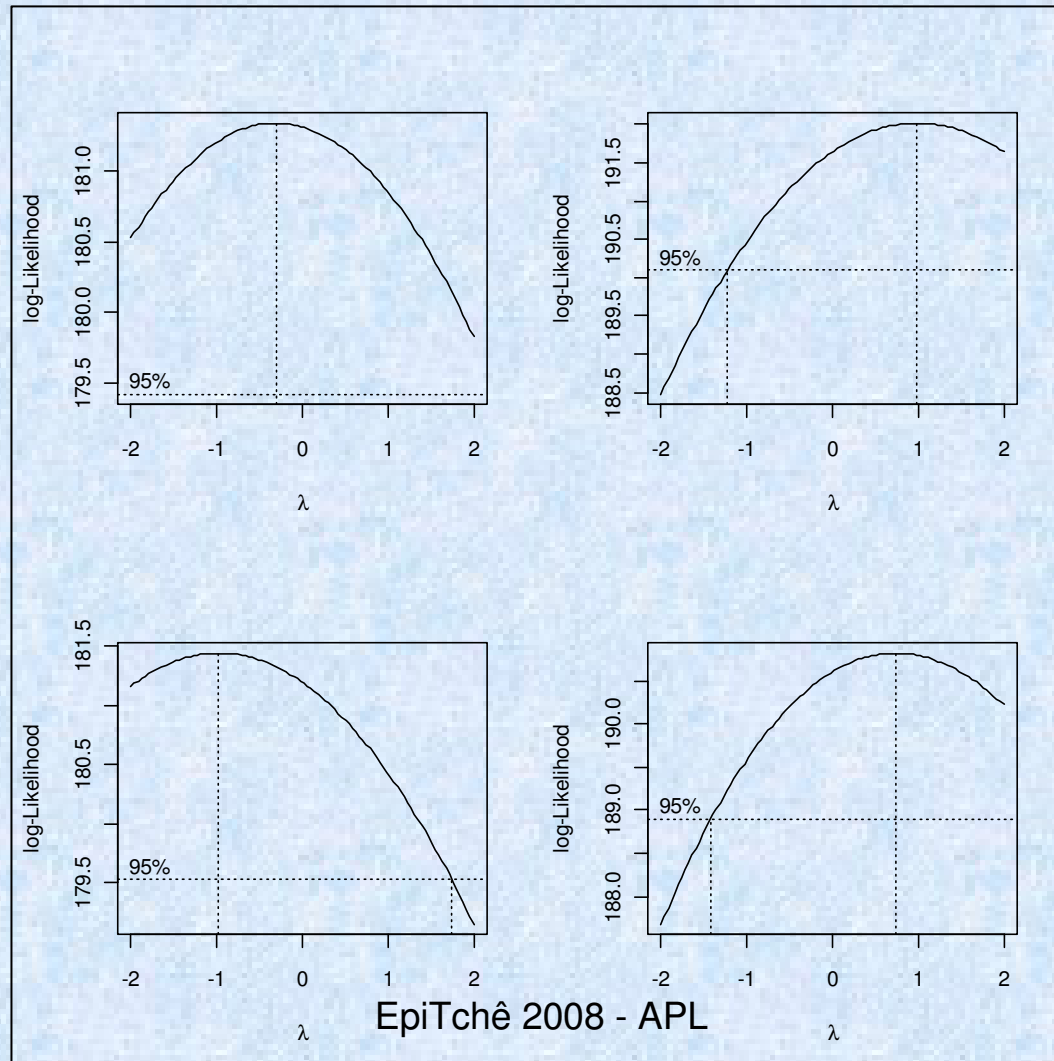
```
boxcox(smr_f ~ rendamd + perc_10r + r_40p)
```

Southeast

```
boxcox(smr_m ~ rendamd + rper_1q + int_pobr + npr)
```

```
boxcox(smr_f ~ rendamd + rper_1q + npr)
```

Box-Cox transformations, based on given sets of independent covariates



Model for men in the Southeast (inverse)

Call:

```
lm(formula = smr_mas ~ rendamd + rper_1q + int_pobr + npr, weights =  
pop_3a_m)
```

Residuals:

Min	1Q	Median	3Q	Max
-28578	-5161	-2193	4174	33282

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	169.55906	15.29949	11.083	4.75e-16	***
rendamd	-0.01219	0.03948	-0.309	0.75857	
rper_1q	0.44979	0.16625	2.706	0.00890	**
int_pobr	-1.23255	0.46118	-2.673	0.00972	**
npr	56.20893	23.15961	2.427	0.01830	*

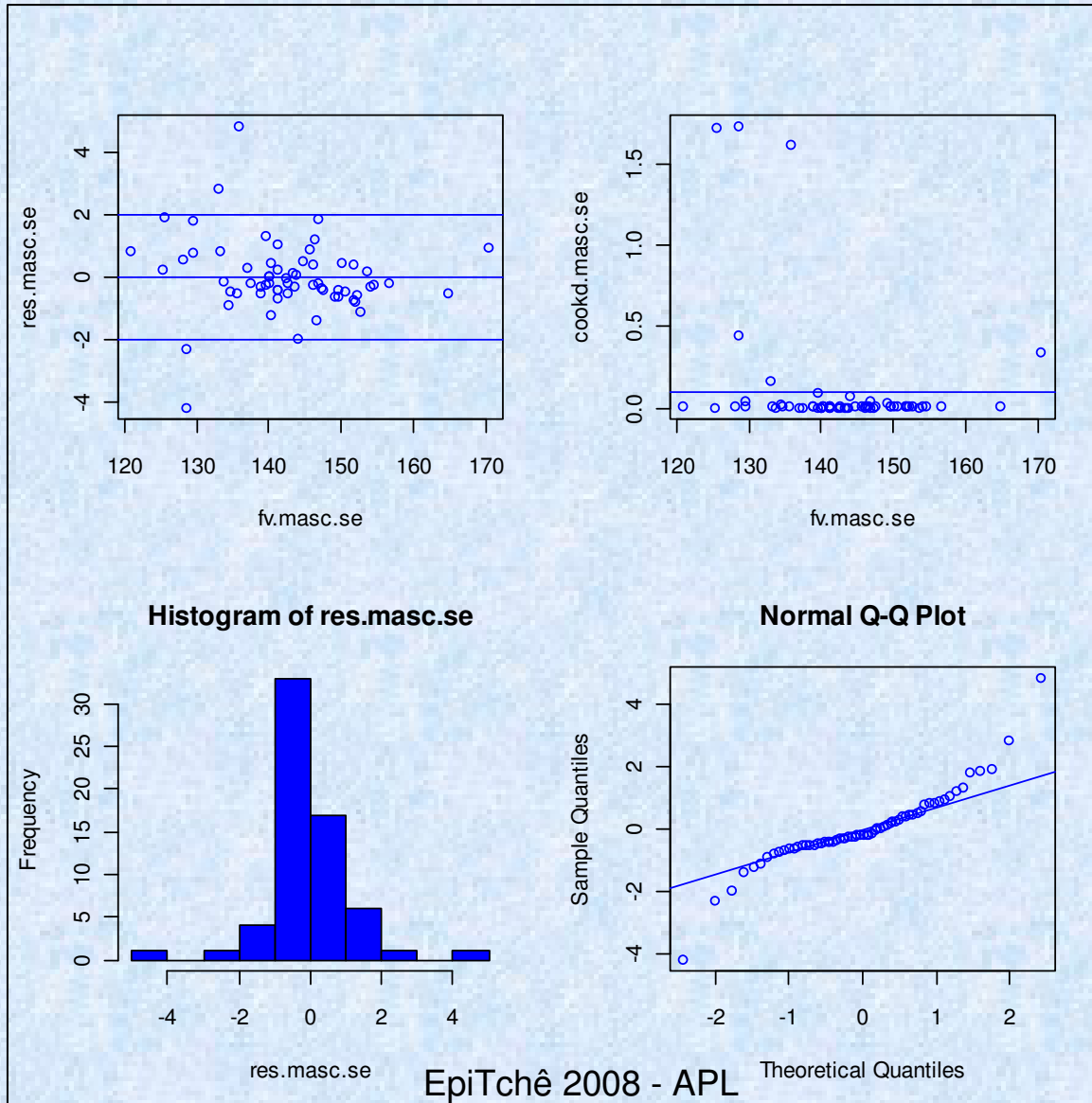
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9876 on 59 degrees of freedom

Multiple R-squared: 0.4682, Adjusted R-squared: 0.4321

F-statistic: 12.98 on 4 and 59 DF, p-value: 1.205e-07

Residual diagnostics, all observations



Model for men in the Southeast, no outliers

Call:

```
lm(formula = smr_mas ~ rendamd + rper_1q + int_pobr + npr, weights =  
pop_3a_m)
```

Residuals:

Min	1Q	Median	3Q	Max
-15805	-3974	-1265	3122	20997

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	115.07887	18.19420	6.325	4.78e-08	***
rendamd	-0.10156	0.04541	-2.237	0.029387	*
rper_1q	0.77592	0.22288	3.481	0.000985	***
int_pobr	0.46935	0.51196	0.917	0.363267	
npr	36.07065	17.23314	2.093	0.040966	*

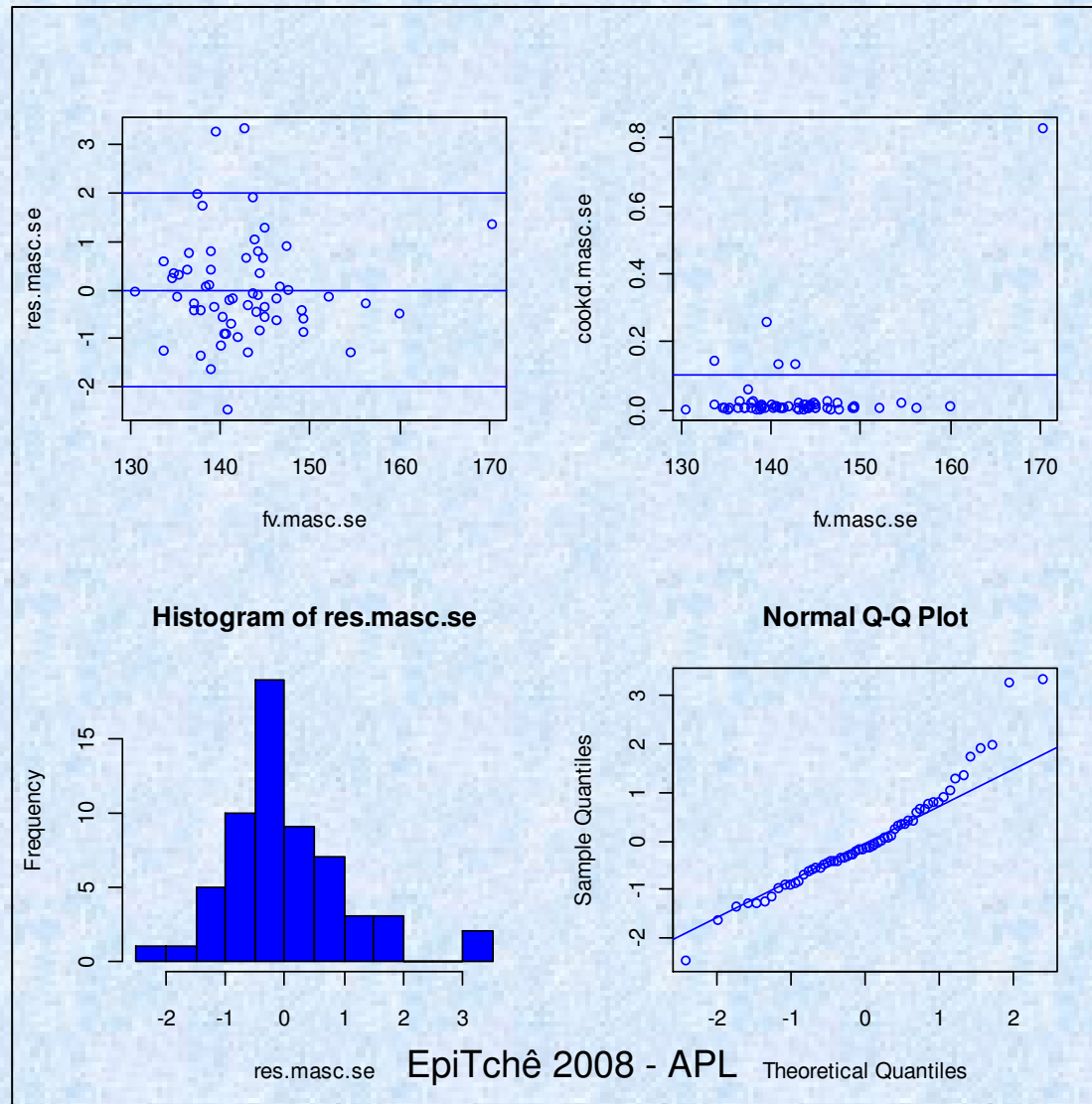
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7135 on 55 degrees of freedom

Multiple R-squared: 0.2354, Adjusted R-squared: 0.1798

F-statistic: 4.234 on 4 and 55 DF, p-value: 0.004654

Residual diagnostics, no outliers (those for Cook's distance > 0.5)



Model for women in the Southeast (identity)

Call:

```
lm(formula = smr_fem ~ rendamd + rper_1q + npr, weights = pop_3a_f)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.28557	-0.08336	0.03916	0.13751	1.00279

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.634e-03	2.998e-04	18.792	< 2e-16 ***
rendamd	3.362e-07	6.337e-07	0.530	0.597768
rper_1q	-1.624e-05	4.093e-06	-3.967	0.000197 ***
npr	-1.306e-03	6.783e-04	-1.925	0.058998 .

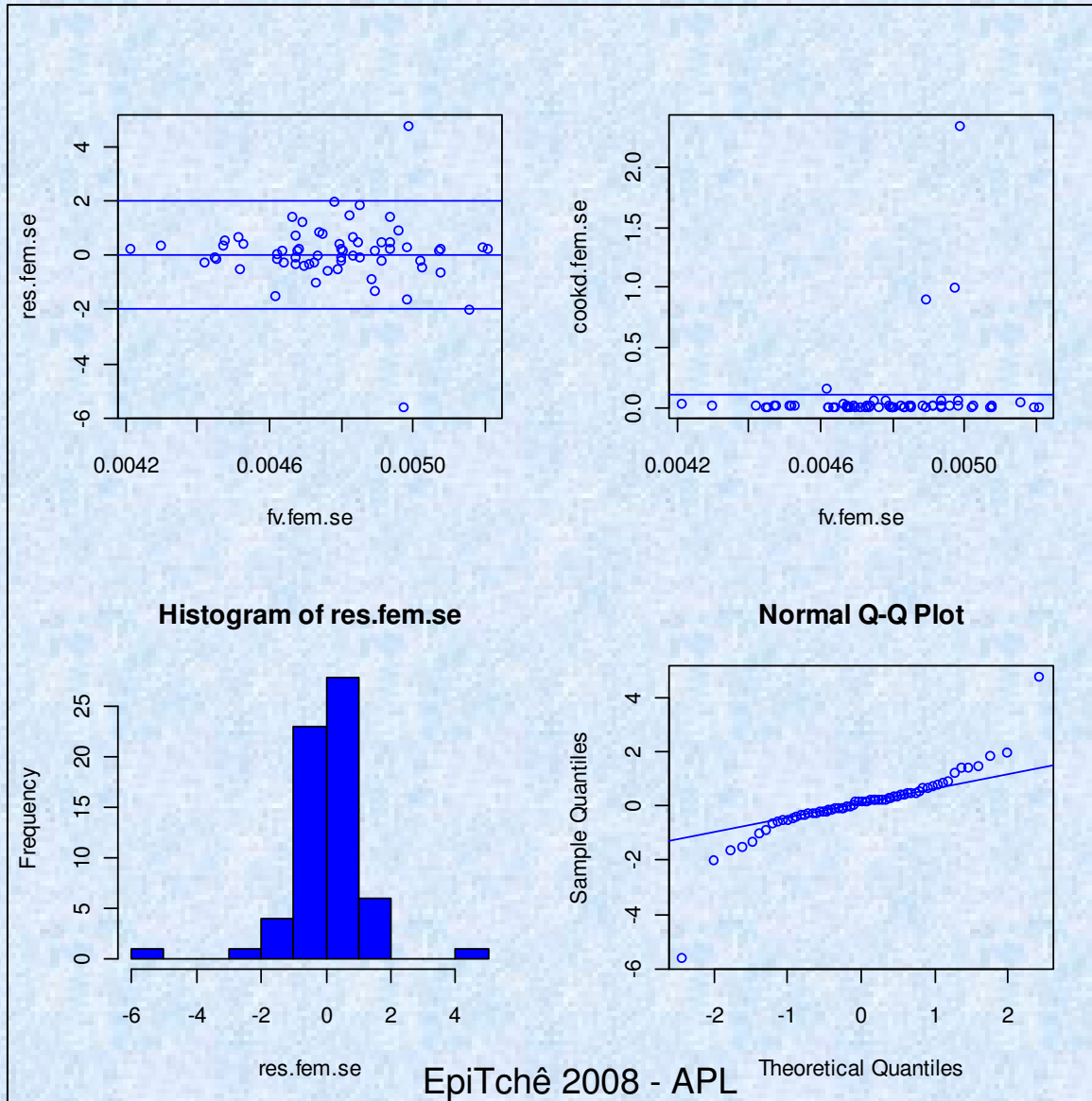
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3069 on 60 degrees of freedom

Multiple R-squared: 0.2214, Adjusted R-squared: 0.1825

F-statistic: 5.689 on 3 and 60 DF, p-value: 0.0017

Residual diagnostics, all observations



Model for women in the Southeast, no outliers

Call:

```
lm(formula = smr_fem ~ rendamd + rper_1q + npr, weights = pop_3a_f)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.60727	-0.10373	0.02134	0.11645	0.59313

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.757e-03	2.537e-04	22.696	<2e-16	***
rendamd	-7.523e-07	9.676e-07	-0.778	0.4401	
rper_1q	-1.078e-05	5.538e-06	-1.947	0.0564	.
npr	-1.282e-03	5.155e-04	-2.488	0.0158	*

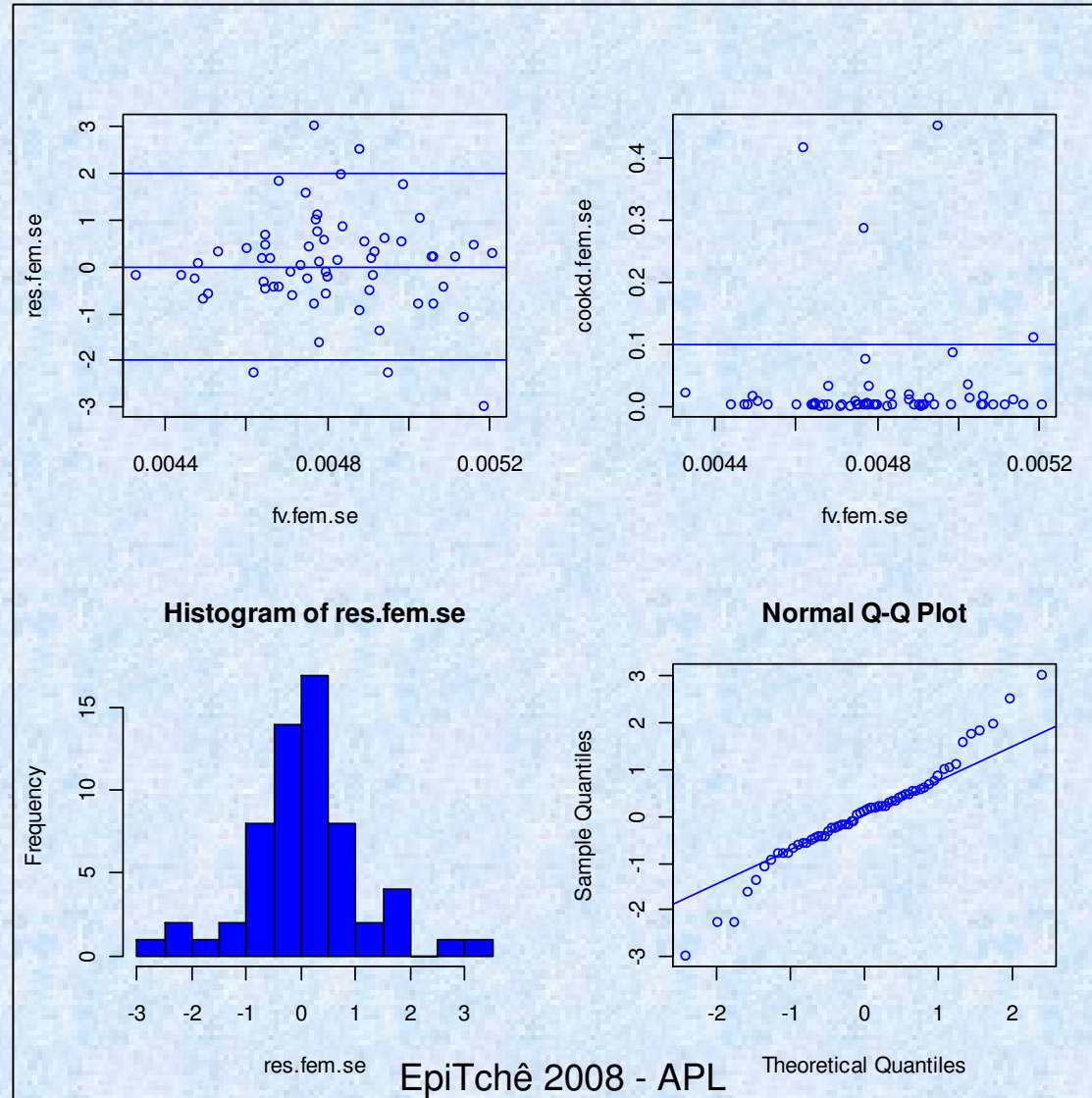
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.224 on 57 degrees of freedom

Multiple R-squared: 0.2346, Adjusted R-squared: 0.1943

F-statistic: 5.822 on 3 and 57 DF, p-value: 0.001529

Residual diagnostics, no outliers (those for Cook's distance > 0.5)



Discussion

- Conglomerate level of analysis is adequate?
- Income from head of household?
- Need of transformations to approach normality;
- Municipalities / metropolitan regions;
- Regression techniques, and specially residual diagnostics as tools for refining and understanding the associations;
- After all, is there an association between income segregation and health ?