Remedies-For-Errors.R

Shraddha Somani

```r
require(faraway)

## Loading required package: faraway

head(uswages)

##          wage educ exper race smsa ne mw so we pt
## 6085  771.60   18    18    0    1  1  0  0  0  0
## 23701 617.28   15    20    0    1  0  0  0  1  0
## 16208 957.83   16     9    0    1  0  0  1  0  0
## 2720  617.28   12    24    0    1  1  0  0  0  0
## 9723  902.18   14    12    0    1  0  1  0  0  0
## 22239 299.15   12    33    0    1  0  0  0  1  0

uswages$exper[uswages$exper < 0] = NA
uswages$race = factor(uswages$race)
levels(uswages$race) = c("White","Black")
uswages$smsa = factor(uswages$smsa)
levels(uswages$smsa) = c("No","Yes")
uswages$pt = factor(uswages$pt)
levels(uswages$pt) = c("No","Yes")
uswages = data.frame(uswages,
                     region =
                       1*uswages$ne +
                       2*uswages$mw +
                       3*uswages$so +
                       4*uswages$we)
uswages$region = factor(uswages$region)
levels(uswages$region) = c("ne","mw","so","we")
uswages = na.omit(uswages)
summary(uswages)

##       wage              educ            exper            race        smsa
##  Min.   :  50.39   Min.   : 0.00   Min.   : 0.00   White:1812   No : 483
##  1st Qu.: 314.69   1st Qu.:12.00   1st Qu.: 8.00   Black: 155   Yes:1484
##  Median : 522.32   Median :12.00   Median :16.00
##  Mean   : 613.99   Mean   :13.08   Mean   :18.74
##  3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00
##  Max.   :7716.05   Max.   :18.00   Max.   :59.00
##        ne               mw               so               we
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000
##  Mean   :0.2278   Mean   :0.2481   Mean   :0.3132   Mean   :0.211
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.000
```

```
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000
##    pt       region
##  No :1802   ne:448
##  Yes: 165   mw:488
##             so:616
##             we:415
##
##
```

```r
# Compute OLS fit to model log(wage)~.
# Perform the Shapiro-Wilk Test of Normality for the residuals, what is the conclusion?
require(car)
```

```
## Loading required package: car

##
## Attaching package: 'car'

## The following objects are masked from 'package:faraway':
##
##     logit, vif
```

```r
m = lm(log(wage) ~ ., uswages)
summary(m)
```

```
##
## Call:
## lm(formula = log(wage) ~ ., data = uswages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5158 -0.3309  0.0504  0.3520  3.9446
##
## Coefficients: (4 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.756885   0.074240  64.074  < 2e-16 ***
## educ         0.087515   0.004494  19.472  < 2e-16 ***
## exper        0.015483   0.001016  15.243  < 2e-16 ***
## raceBlack   -0.215273   0.048723  -4.418 1.05e-05 ***
## smsaYes      0.177741   0.030177   5.890 4.54e-09 ***
## ne          -0.046846   0.038922  -1.204    0.229
## mw          -0.038526   0.038021  -1.013    0.311
## so          -0.036765   0.036635  -1.004    0.316
## we                 NA         NA      NA       NA
## ptYes       -1.067418   0.046344 -23.032  < 2e-16 ***
## regionmw           NA         NA      NA       NA
## regionso           NA         NA      NA       NA
## regionwe           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.5686 on 1958 degrees of freedom
## Multiple R-squared:  0.368,  Adjusted R-squared:  0.3654
## F-statistic: 142.5 on 8 and 1958 DF,  p-value: < 2.2e-16
```

```
shapiro.test(residuals(m))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(m)
## W = 0.96353, p-value < 2.2e-16
```

```
# Answer
# - The null hypothesis is that the the residuals are normal.
# - Since the p-value is smaller than the significant value (0.05), we reject
the null hypothesis.
# - Hence, we can say that the residuals are not normal.


# Compute WLS fit to model log(wage)~. and weights = 1/(1+educ)
# Perform the Shapiro-Wilk Test of Normality for the residuals, what is the
conclusion?
m1 = lm(log(wage) ~ ., uswages, weight = 1/(1 + educ))
shapiro.test(residuals(m1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(m1)
## W = 0.972, p-value < 2.2e-16
```

```
# Answer
# - The null hypothesis is that the the residuals are normal.
# - Since the p-value is smaller than the significant value (0.05), we reject
the null hypothesis.
# - Hence, we can say that the residuals are not normal.


# Compute Robust fit to model log(wage)~. using Huber, Hampel, Biquare, LTS,
and LAD
# Compare coefficients of the above fits using OLS, WLS, Huber, Hampel,
Biquare, LTS, and LAD
# Which would you recommend?
# Why?
require(MASS)
```

```
## Loading required package: MASS
```

```
# Huber M-Estimation
m2 = rlm(log(wage) ~ educ + exper + race + smsa + pt, psi = psi.huber,
```

```r
uswages)
# Hample M-estimation
m3 = rlm(log(wage) ~ educ + exper + race + smsa + pt, psi = psi.hampel, init
= "lts", maxit = 100, uswages)
# Tukey Bisquare M-estimation
m4 = rlm(log(wage) ~ educ + exper + race + smsa + pt, psi = psi.bisquare,
init = "lts", maxit = 100, uswages)
# Least Trimmed Squares (LTS)
require(robustbase)

## Loading required package: robustbase

##
## Attaching package: 'robustbase'

## The following object is masked from 'package:faraway':
##
##     epilepsy

m5 = ltsReg(log(wage) ~ educ + exper + race + smsa + pt, data = uswages)

## Warning in covMcd(X, alpha = alpha, use.correction = use.correction): The
986-th order statistic of the absolute deviation of variable 3
## is zero.
## There are 1812 observations (in the entire dataset of 1967 obs.)
## lying on the hyperplane with equation a_1*(x_i1 - m_1) + ... +
## a_p*(x_ip - m_p) = 0 with (m_1, ..., m_p) the mean of these
## observations and coefficients a_i from the vector a <- c(0, 0, 1,
## 0, 0)

m6 = ltsReg(log(wage) ~ educ + exper + race + smsa + pt, data = uswages,
nsamp = "exact")

## Warning in .fastlts(x, y, h, nsamp, intercept, adjust, trace =
as.integer(trace)): 'nsamp' options 'best' and 'exact' not allowed for n
greater than 599. Will use default.

## Warning in .fastlts(x, y, h, nsamp, intercept, adjust, trace =
as.integer(trace)): The 986-th order statistic of the absolute deviation of
variable 3
## is zero.
## There are 1812 observations (in the entire dataset of 1967 obs.)
## lying on the hyperplane with equation a_1*(x_i1 - m_1) + ... +
## a_p*(x_ip - m_p) = 0 with (m_1, ..., m_p) the mean of these
## observations and coefficients a_i from the vector a <- c(0, 0, 1,
## 0, 0)

# Least Absolution Deviation (LAD)
require(quantreg)

## Loading required package: quantreg
```

```
## Loading required package: SparseM

##
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':
##
##     backsolve

m7 = rq(log(wage) ~ educ + exper + race + smsa + pt, data = uswages)
# Comparing Coefficients
coefs <- compareCoefs(m, m2, m3, m4, m5, m6, m7, se = FALSE)

## Warning in compareCoefs(m, m2, m3, m4, m5, m6, m7, se = FALSE): models to
## be compared are of different classes

##
## Call:
## 1: lm(formula = log(wage) ~ ., data = uswages)
## 2: rlm(formula = log(wage) ~ educ + exper + race + smsa + pt, data =
##   uswages, psi = psi.huber)
## 3: rlm(formula = log(wage) ~ educ + exper + race + smsa + pt, data =
##   uswages, psi = psi.hampel, init = "lts", maxit = 100)
## 4: rlm(formula = log(wage) ~ educ + exper + race + smsa + pt, data =
##   uswages, psi = psi.bisquare, init = "lts", maxit = 100)
## 5: ltsReg.formula(formula = log(wage) ~ educ + exper + race + smsa +
##   pt, data = uswages)
## 6: ltsReg.formula(formula = log(wage) ~ educ + exper + race + smsa +
##   pt, data = uswages, nsamp = "exact")
## 7: rq(formula = log(wage) ~ educ + exper + race + smsa + pt, data =
##   uswages)
##              Est. 1  Est. 2  Est. 3  Est. 4  Est. 5  Est. 6  Est. 7
## (Intercept)  4.7569  4.6335  4.6172  4.5857                  4.6396
## educ         0.0875  0.0961  0.0966  0.0997  0.1000  0.0999  0.0963
## exper        0.0155  0.0162  0.0160  0.0167  0.0174  0.0174  0.0167
## raceBlack   -0.2153 -0.2069 -0.2097 -0.2131 -0.2374 -0.2383 -0.2599
## smsaYes      0.1777  0.1611  0.1631  0.1594  0.1640  0.1650  0.1833
## ne          -0.0468
## mw          -0.0385
## so          -0.0368
## we
## ptYes       -1.0674 -1.1494 -1.1401 -1.1694 -1.2084 -1.2014 -1.1826
## regionmw
## regionso
## regionwe
## Intercept                                    4.5887  4.5899

colnames(coefs) <- c("OLS", "Huber", "Bisquare", "Hample", "LTS-
exact", "LAD")
coefs
```

```
##                     OLS       Huber    Bisquare      Hample        LTS
## (Intercept)  4.75688510  4.63347176  4.61716382  4.58567071         NA
## educ         0.08751484  0.09614668  0.09663798  0.09967062  0.1000266
## exper        0.01548319  0.01618256  0.01604057  0.01668646  0.0174227
## raceBlack   -0.21527343 -0.20692145 -0.20970829 -0.21311066 -0.2374227
## smsaYes      0.17774071  0.16107905  0.16310193  0.15944017  0.1640238
## ne          -0.04684583          NA          NA          NA         NA
## mw          -0.03852593          NA          NA          NA         NA
## so          -0.03676539          NA          NA          NA         NA
## we                   NA          NA          NA          NA         NA
## ptYes       -1.06741757 -1.14937568 -1.14008600 -1.16944555 -1.2084399
## regionmw             NA          NA          NA          NA         NA
## regionso             NA          NA          NA          NA         NA
## regionwe             NA          NA          NA          NA         NA
## Intercept            NA          NA          NA          NA  4.5886638
##               LTS-exact         LAD
## (Intercept)          NA  4.63955830
## educ         0.09994071  0.09630694
## exper        0.01738042  0.01666800
## raceBlack   -0.23829355 -0.25991098
## smsaYes      0.16502113  0.18334795
## ne                   NA          NA
## mw                   NA          NA
## so                   NA          NA
## we                   NA          NA
## ptYes       -1.20136371 -1.18258840
## regionmw             NA          NA
## regionso             NA          NA
## regionwe             NA          NA
## Intercept    4.58989314          NA

# Answer:
# - We see that LTS and LTS-exact appear to agree with each other and both
are very different from OLS.
# - All three M-estimation methods, Huber, Bisquare, and Hample are different
from each other, and different from OLS and both LTS's.
# - LAD is similar to OLS.
# - LTS is recommended since it has the best breakdown.
```