EDA Oct 5

ex2

a)

First fit from the pair plot we can see RESP and PVRTY have the strongest linear relationship.

fit by PVRTY: RESP = 74.6773926 + 1.8493319*PVRTY + RESP.1

Second fit sweep PVRTY out of all the other variables, and plot their scatter plots against RESP.1. From observation, choose INFM as the second carrier.

INFM sweep out of PVRTY: INFM = 98.9991817 + 0.9492872*PVRTY + INFM.1

fit RESP.1 by INFM.1: RESP.1 = -1.9064114 + -0.21438*INFM.1 + RESP.14

Third fit sweep INFM.1 out of all the other variables.1, and plot their scatter plots against RESP.14. From observation, choose SING as the third carrier.

PVRTY sweep out of SING: SING = 31.4965064 + 0.1200699*PVRTY + SING.1

INFM.1 sweep out of SING.1: SING.1 = 0.0122708 + -0.0419573*INFM.1 + SING.14

fit RESP.14 by SING.14: RESP.14 = -0.0959464 + 0.8789384*SING.14 + RESP.142

result RESP = 65.74 + 1.57PVRTY - 0.21INFM + 0.88SING

b)

The outlier does not affect my my selection of variables too much.

c)

```r
lm(RESP.t~PVRTY.t+SING.t+RINC.t+INFM.t+WARD.t)##least square without outlier
```

```
##
## Call:
## lm(formula = RESP.t ~ PVRTY.t + SING.t + RINC.t + INFM.t + WARD.t)
##
## Coefficients:
## (Intercept)      PVRTY.t       SING.t       RINC.t       INFM.t
##     49.6068       2.0406       3.8048      -0.6802      -0.4508
##      WARD.t
##      0.5807
```

```r
lm(RESP~PVRTY+SING+RINC+INFM+WARD,data = ex1)##least square with outlier
```

```
##
## Call:
## lm(formula = RESP ~ PVRTY + SING + RINC + INFM + WARD, data = ex1)
##
## Coefficients:
## (Intercept)        PVRTY         SING         RINC         INFM
##     63.5423       1.6353       2.2843      -0.2891      -0.4060
##        WARD
##      0.8200
```

from the comparison, we can see least square regression is more effected by the outlier, with intercept and coefficent for RINC, INFM and WARD increasing, coefficient for PVRTY and SING decreasing.

ex4

a)

sum of squard residuals advantage:we have a close form for the fit that minimize the sum of squard residuals. disadvantage:least square regression line is always the best under this criterion, while LSR is unrobust to outliers.

sum of absolute residuals advantage:very intuitive, measures the distance between fit and real data. disadvantage:there's no close form to minimize the sum of absolute residuals, have to do iteration. and there can be multiple lines that minimize it. unrobust to outliers.

fourth spread of residuals advantage:easy to calculate compared to other criteria. robust to outliers disadvantage:fourth-spread is robust, with the trade-off of sensitivity. two set of residuals can differ considerably while fourth-spreads stay close.

b)

sum of absobulte residuals is the most suitable criterion.

R^2 statistic defined as $\frac{SSR}{SSTO}$ in linear regression analysis. it measures how much total variance is captured in the regression fit.

c)

when we already know the data come from multivariate normal distribution, which meets the error term assumption in LSR, we benefit from the close form trait of least square. least square regression theory provides well-developed inference without worrying whether the error terms follow normal distribution.

ex5

a)

```r
medx<-medpolish(x)
```

```
## 1: 302
## Final: 300
```

```r
medy<-medpolish(y)
```

```
## 1: 362
## Final: 362
```

```r
ex<-medx$residuals
ey<-medy$residuals
```

```
##          a        b    |res|
##  1 -0.81955  8.55639 2367.988
##  2 -0.52616  1.54989 2769.802
##  3 -0.35421 -0.16240 2725.792
##  4  0.18270  0.05128 2739.690
##  5 -0.05769 -0.01619 2735.301
##  6  0.01822  0.00511 2736.687
```

```
##  7 -0.00575 -0.00162 2736.249
##  8  0.00182  0.00051 2736.387
##  9 -0.00057 -0.00016 2736.344
## 10  0.00018  0.00005 2736.358
##    -1.56103  9.98287 2736.358
```

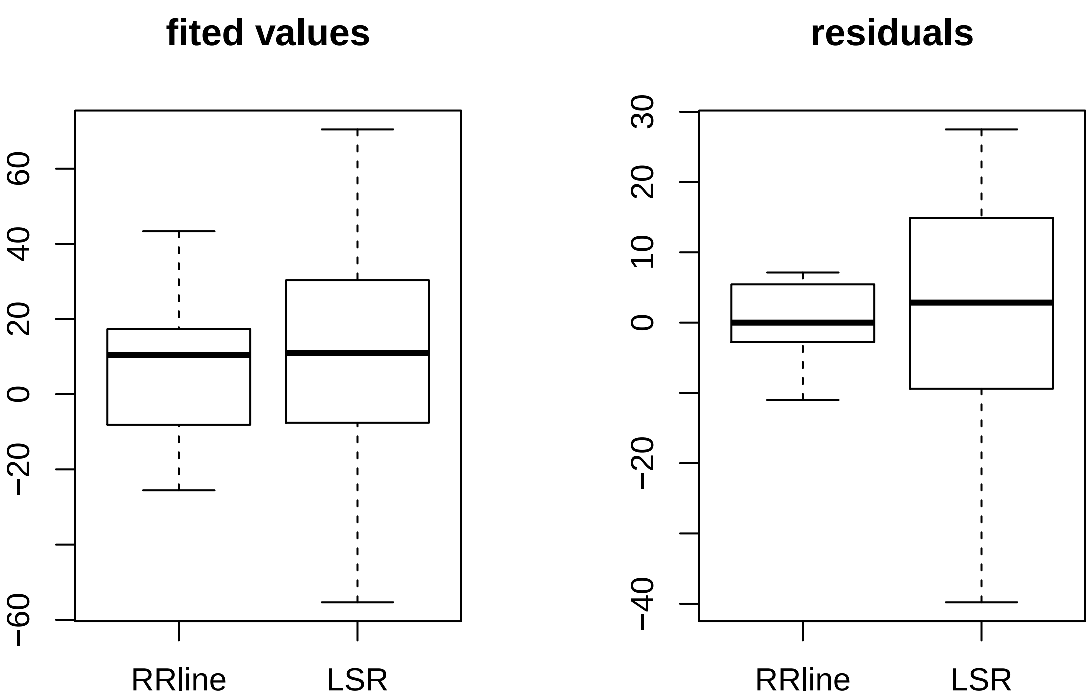
In comparison to the result of data without outliers, change in coefficients is slight. ####b)

```
##
## Call:
## lm(formula = as.vector(ey) ~ as.vector(ex))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.541  -7.981   2.851  14.824  27.488
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.00398    4.17976  -0.958    0.349
## as.vector(ex)  0.10916    0.07814   1.397    0.176
##
## Residual standard error: 20.08 on 22 degrees of freedom
## Multiple R-squared:  0.08149,    Adjusted R-squared:  0.03974
## F-statistic: 1.952 on 1 and 22 DF,  p-value: 0.1763
```

The coefficients differ a lot from table 7-8.

c)

boxplots



From the boxplots we can clearly see using RRline, both fited values and residuals have smaller variance. LSR method tries to evenly distribute data on both side of the line, causing the boxplot to be more symmetric. In comparison to Figure 7-10, outliers apparently affect LSR more than RRline method, due to unrobust nature of LSR.