EDA homework 6

1. *Freeny data*: In R, type: `data("freeny")`. These are the data used in the midterm problem 7: Y = Column 1 (Quarterly revenue); X1, X2, X3 are Columns 3,4,5 (Price Income, Income Level, Market Potential), respectively.

(a) Fit a least squares model to these data:

```
lm.freeny <- lm(freeny[,1] ~ freeny[,3] + freeny[,4] + freeny[,5], data=freeny)
```

the formula:

y = -13.3101438 + -0.8348827X1 + 0.8455586X2 + 1.6273453X3

(b) Now fit a robust fit by sweeping out, in turn, X3, then X1, then X2.

the formula: y = -30.72755 + 3.0424095X3 + -0.4998425X1 + 0.4190509X2

(c) Now fit a robust fit by sweeping out, in turn, X1, then X3, then X2.

the formula: y = -30.50599 + -0.4884994X1 + 2.9971882X3 + 0.471577X2

(d) Compare the three sets of coefficients. What do you observe?

The coefficients from linear fit differs a lot from (b)&(c), while RRfit produced similar results. Interpretation: outliers exert serious distortion on the coefficients in linear regression, RRfit is robust, the outcome of different sequence in fit change only a little bit.

2. *Smoking data*: Below are the smoking rates (in percentages, times 10) for 9 different years, for 4 race x gender groups and 4 categories of education level (< 12 years of schools, HS graduate, some college, college graduate):

(a) Fit RRline to each of the 16 rows (1 for each race x gender x education level). Use centercept at 1990 (we don't care about smoking rates at year 0).

White male:

```
y = 423.1001157 + -5.8062428x

y = 330 + -5.7409819x

y = 268.6666667 + -8.6666667x

y = 154.8786651 + -7.8786651x
```

Black male:

```
y = 453.6154192 + -4.3077096x

y = 401.9231369 + -7.4615684x

y = 292.3541667 + -10.5486111x

y = 206 + -15.2595205x
```

White female

```
y = 339.090973 + -5.8062428x

y = 280.4513809 + -2.7097238x

y = 229.9166667 + -6.9166667x
```

```
  y = 140.1666667 + -7.1666667x
```

Black female

```
  y = 335.3021155 + -1.8489423x
```

```
  y = 270.1259645 + -9.3046272x
```

```
  y = 251.3333333 + -3.3333333x
```

```
  y = 169.3390842 + -16.8087023x
```

(b) Place the centercepts in a 4x4 table (rows = education, columns = race-gender group), and perform median polish. Calculate residuals.

```
## 1: 252.5245
## Final: 252.5245
```

```
##
## Median Polish Results (Dataset: "twoway")
##
## Overall: 280.0315
##
## Row Effects:
## [1]  19.30184  66.21929 -24.84747 -19.30184
##
## Column Effects:
## [1]   95.63579   27.96701  -27.96701 -127.63407
##
## Residuals:
##         [,1]     [,2]     [,3]    [,4]
## [1,]  28.131   2.6997  -2.6997 -16.821
## [2,]  11.729  27.7053 -25.9296 -12.617
## [3,] -11.729  -2.6997   2.6997  12.617
## [4,] -21.063 -18.5707  18.5707  36.244
```

(c) Place the slopes in a 4x4 table (rows = education, columns = race-gender group), and perform median polish. Calculate residuals.

```
## 1: 29.84618
## 2: 29.28479
## Final: 29.28479
```

```
##
## Median Polish Results (Dataset: "twoway")
##
## Overall: -7.670832
##
## Row Effects:
## [1]  0.4670079 -1.3342577  2.8576369 -0.4670079
##
## Column Effects:
## [1]  3.618929  1.503182 -1.503182 -4.303951
##
## Residuals:
##         [,1]      [,2]      [,3]    [,4]
## [1,] -2.2213 -0.040339  0.040339  3.6291
```

```
## [2,]  1.0785  0.040339 -0.040339 -1.9505
## [3,] -1.0785  0.600290 -0.600290  1.9505
## [4,]  2.6700 -2.669969  6.307689 -4.3669
```

(d) What do the results suggest about the typical smoking rate, by race-gender group, around 1990?

The black communiy had higher smoking rate than their white conuterpart of the same gender and education, famele had lower smoking rate compared to their male counterpart.

(e) What do the results suggest about the trend in smoking rates by race-gender group between 1974 and 1992?

The smoking rate decreases. The higer education level, the more decrease tendency (with one execption black female with HS diploma has a slope of -9.3, much higher than the some-college group).