Regression Review and Robust Regression

Slides prepared by Elizabeth Newton (MIT)

S-Plus Oil City Data Frame

Monthly Excess Returns of Oil City Petroleum, Inc. Stocks and the Market

SUMMARY:

The oilcity data frame has 129 rows and 2 columns. The sample runs from April 1979 to December 1989. This data frame contains the following columns:

VALUE:

Oil

monthly excess returns of Oil City Petroleum, Inc. stocks.

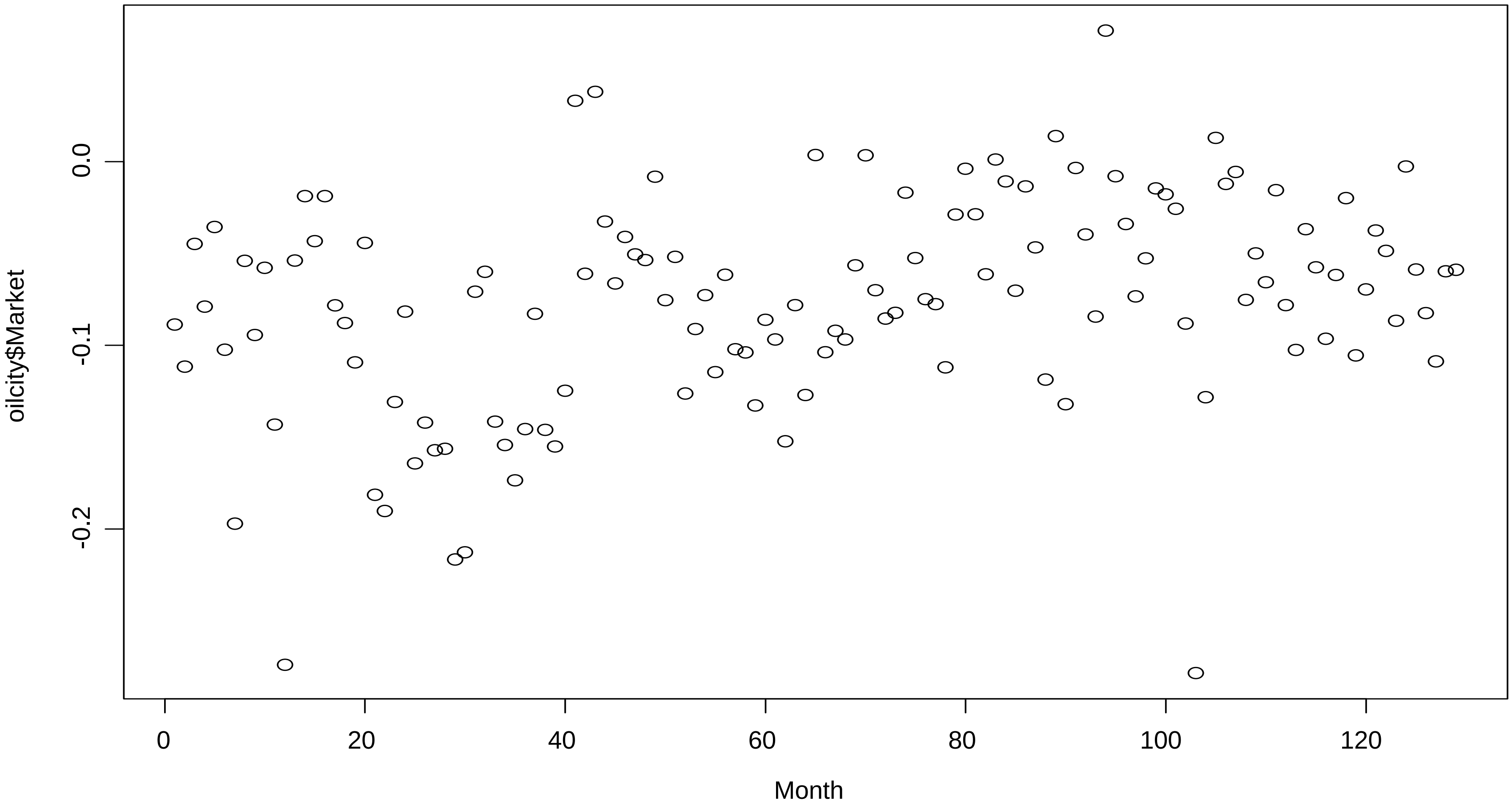
Market

monthly excess returns of the market.

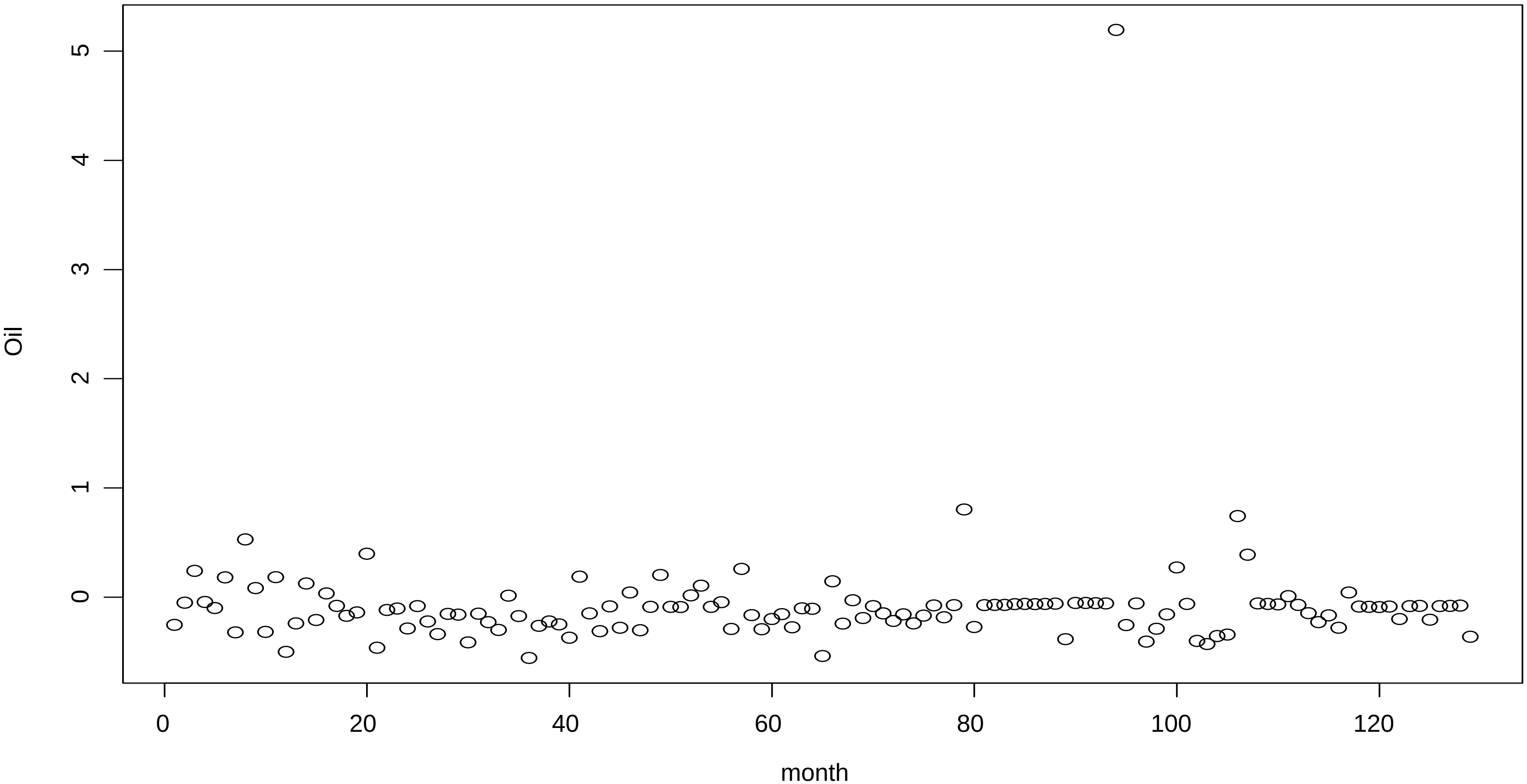
Oil City Data (continued)

- Returns = relative change in the stock price over a one month interval
- Excess returns are computed relative to the monthly return of a 90-day US Treasury bill at the risk-free rate
- Financial economists use least squares to fit a straight line predicting a particular stock return from the market return.
- Beta= estimated coefficient of the market return. Measures the riskiness of the stock in terms of standard deviation and expected returns.
- Large beta -> stock is risky compared to market, but also expected returns from the stock are large.

Plot of Market returns vs. month

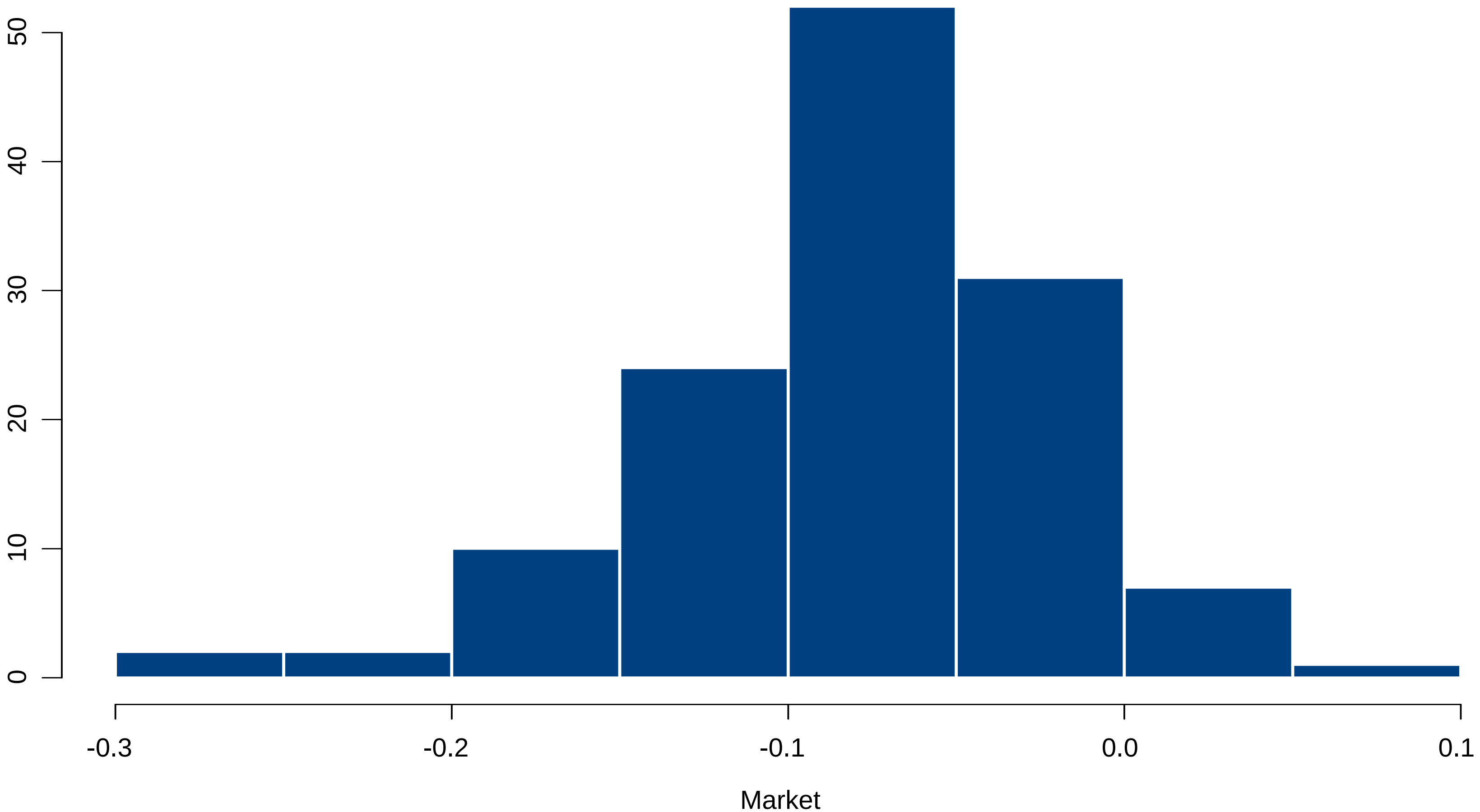


Plot of Oil City Petroleum return vs. month

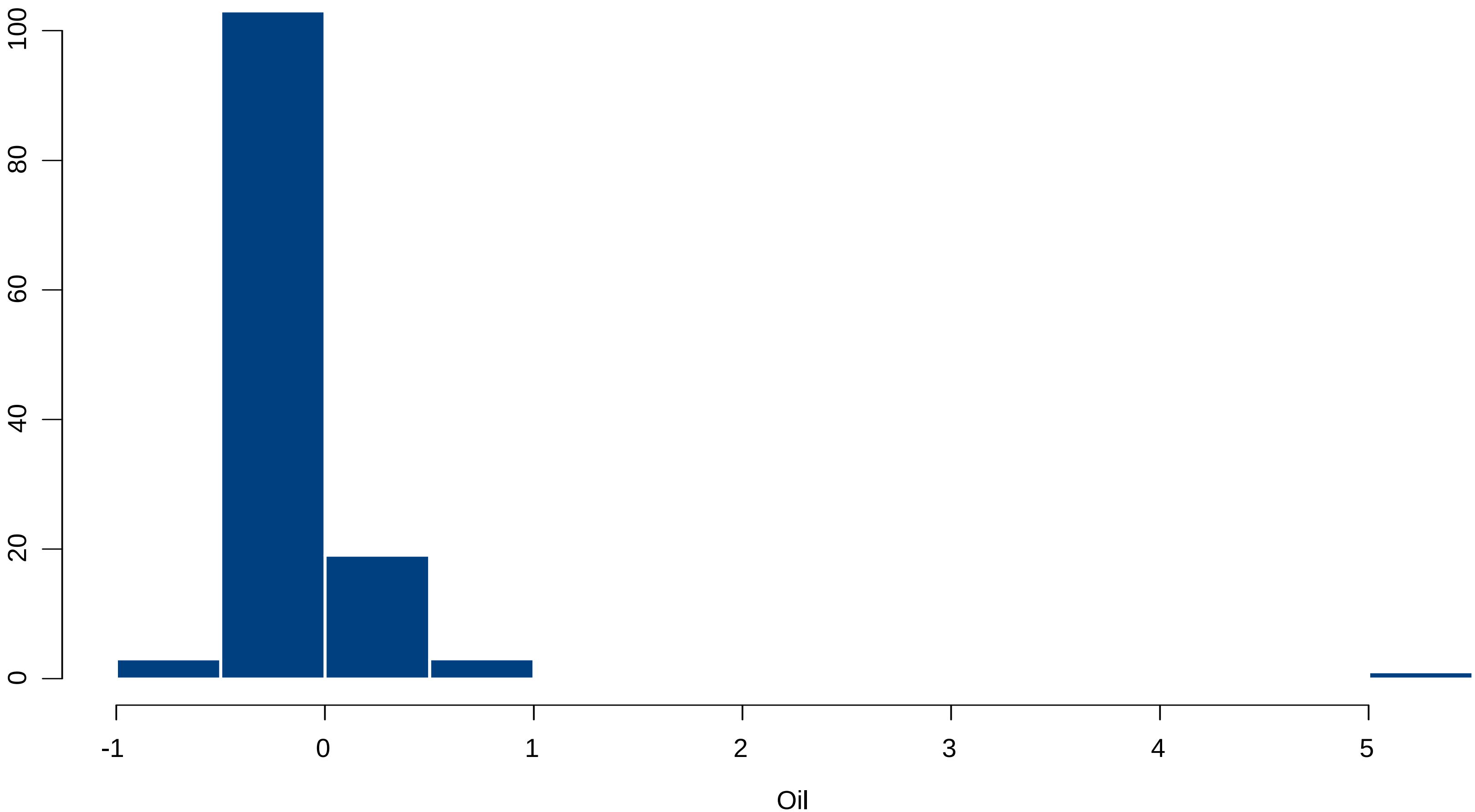


This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

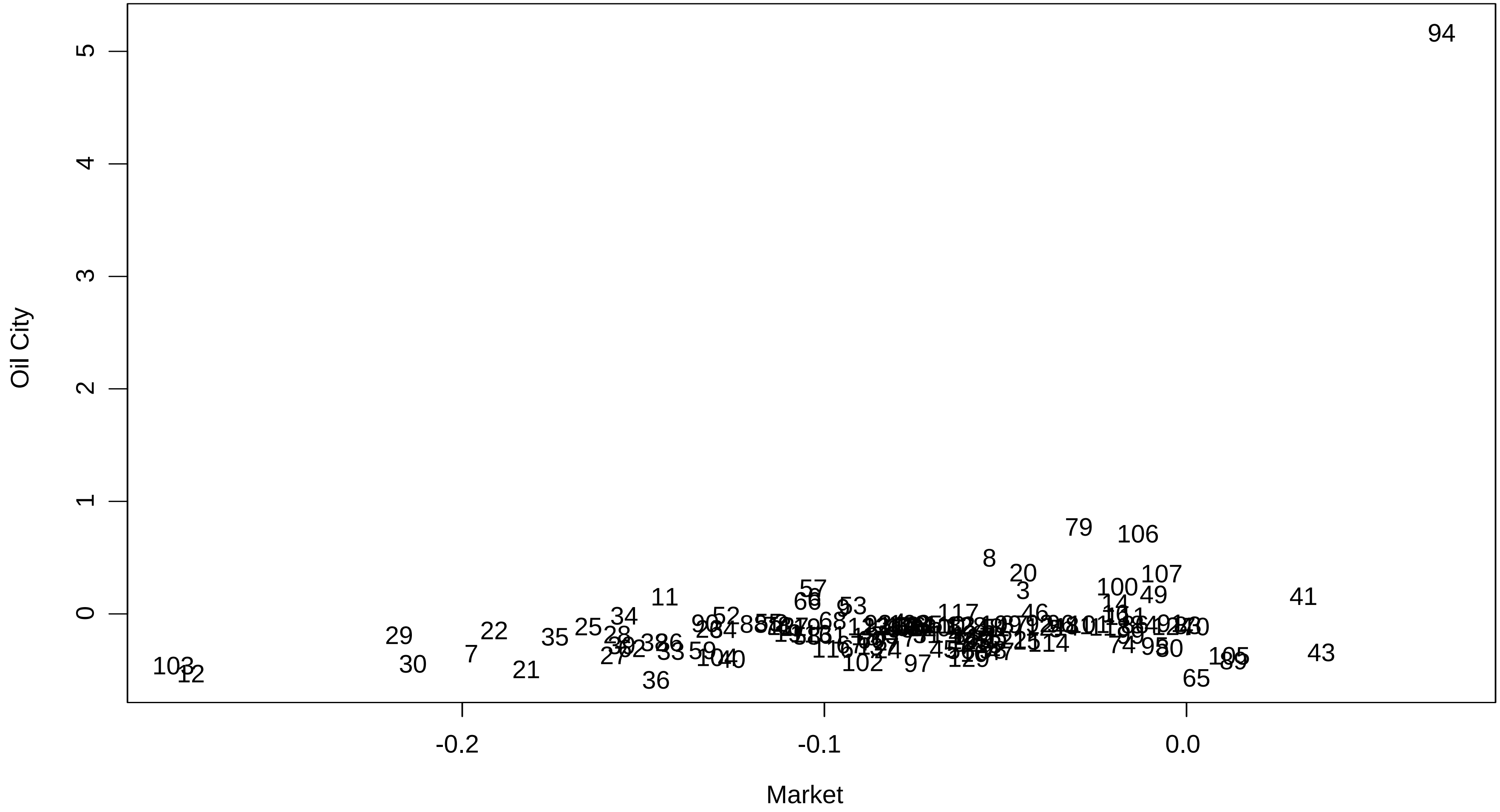
Histogram of Market Returns



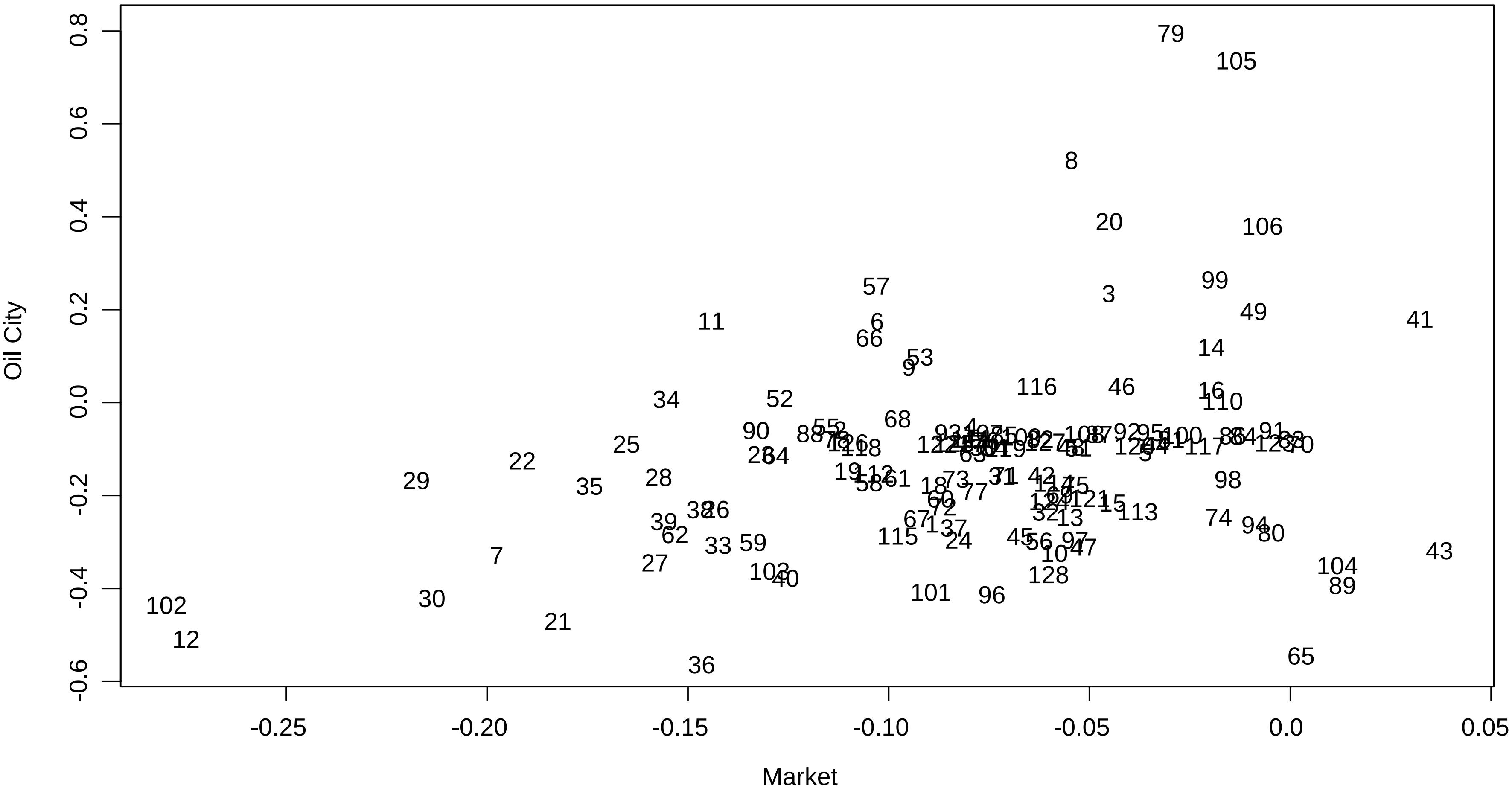
Histogram of Oil City Returns



Plot of Oil City vs. Market Returns



Plot of Oil City vs. Market Returns without observation 94

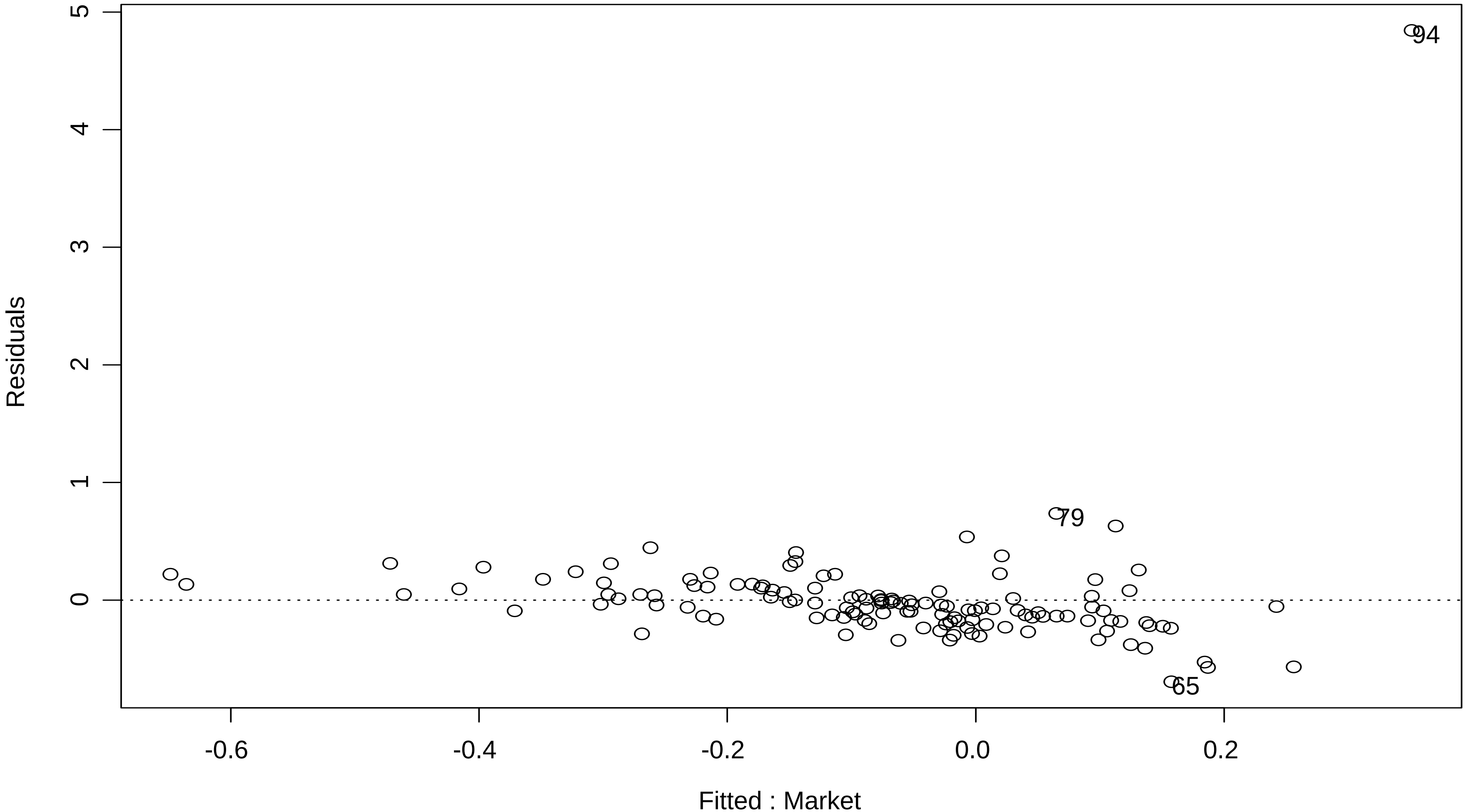


```
> summary(oilcity)
      Oil                       Market
    Min.:-0.55667260        Min.:-0.27857020
 1st Qu.:-0.23968330     1st Qu.:-0.10557534
  Median:-0.10049000      Median:-0.07277544
    Mean:-0.07221215        Mean:-0.07689209
 3rd Qu.:-0.05821000     3rd Qu.:-0.03973828
    Max.: 5.19292000        Max.: 0.07131940
```

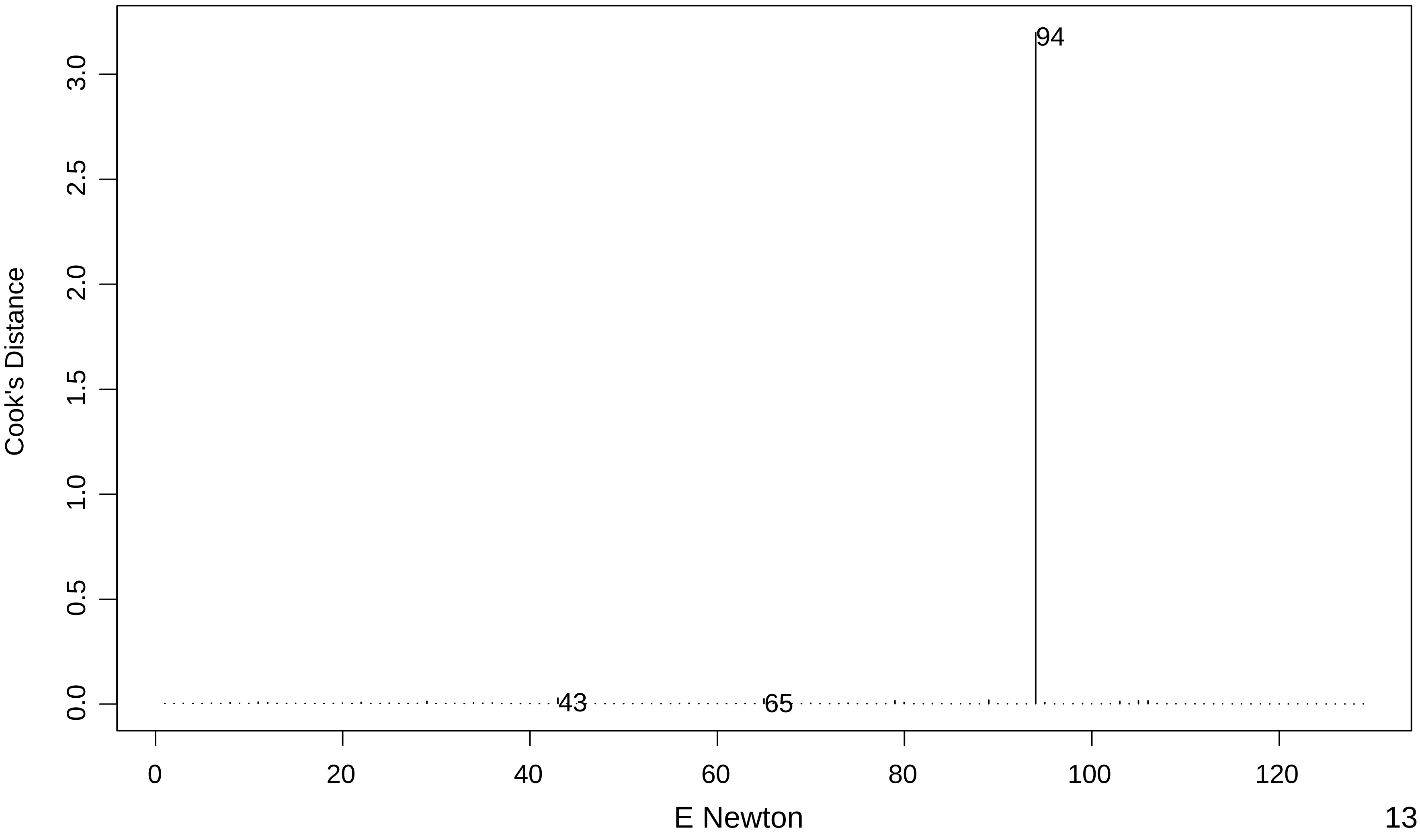
Summary oil.lm

```
Call: lm(formula = Oil ~ Market, data = oilcity)
Residuals:
     Min       1Q   Median       3Q   Max
 -0.6952 -0.1732 -0.05444 0.08407 4.842

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 0.1474 0.0707     2.0849  0.0391
     Market 2.8567 0.7318     3.9040  0.0002

Residual standard error: 0.4867 on 127 degrees of freedom
Multiple R-Squared: 0.1071
F-statistic: 15.24 on 1 and 127 degrees of freedom, the p-value
   is 0.0001528

Correlation of Coefficients:
       (Intercept)
Market 0.7956
```

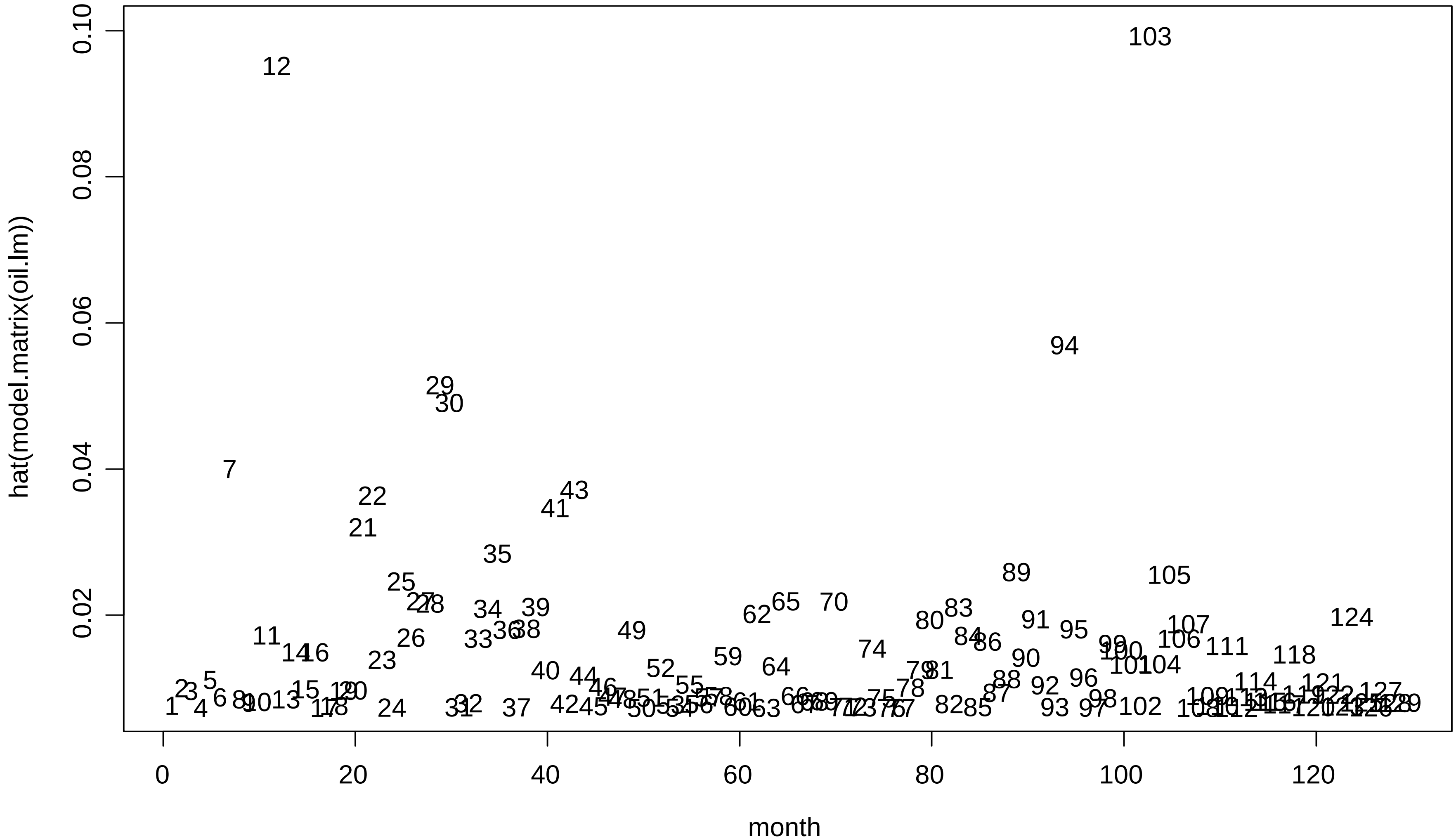
Plot of residual vs. fit for oil.lm



Plot of Cooks Distance vs. Index



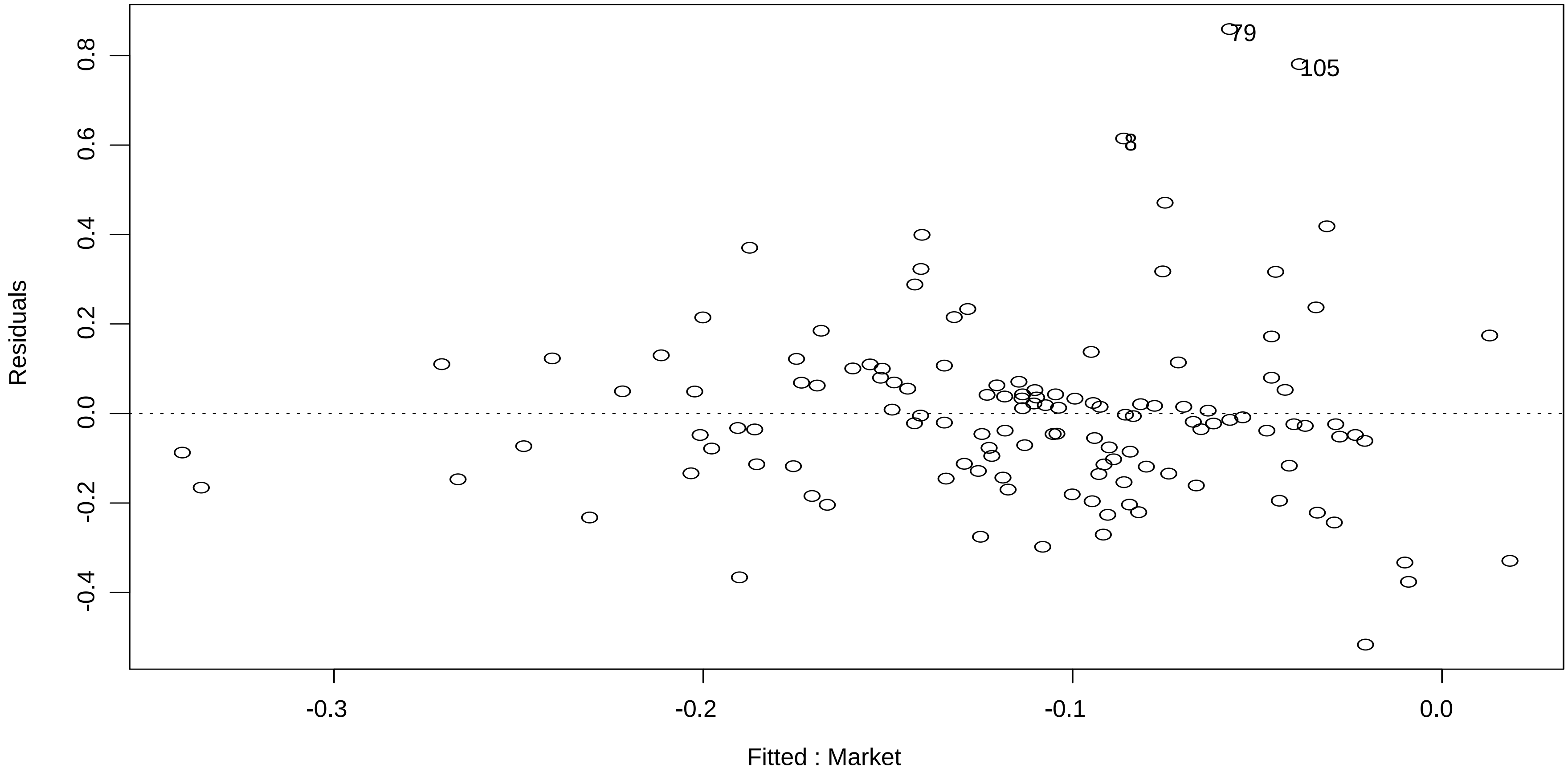
Plot of hat matrix diagonals for oil.lm



Summary of model without observation 94

```
Call: lm(formula = Oil ~ Market, data = oilcity94)

Residuals:
     Min       1Q   Median       3Q   Max
 -0.5169 -0.1174 -0.01959 0.06864 0.859

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) -0.0247  0.0304    -0.8139  0.4173
     Market  1.1355  0.3137     3.6202  0.0004

Residual standard error: 0.2033 on 126 degrees of freedom
Multiple R-Squared: 0.09422
F-statistic: 13.11 on 1 and 126 degrees of freedom, the p-value
   is 0.0004249

Correlation of Coefficients:
       (Intercept)
Market 0.8061
```

Plot of residual vs fit for model without observation 94



Weighted Least Squares

Used when observations, y_i, have unequal variances

$$y = X\beta + \varepsilon$$

$$E(\varepsilon) = 0,\ Var(\varepsilon) = \sigma^2 V$$

V is non-singular positive definite

V is diagonal if errors are uncorrelated,

V is always symmetric

$\exists$ nxn non-singular symmetric matrix, R

such that $R'R = RR = V$

R is sometimes called the square root of V

Weighted least squares (continued)

Define new variables :

$$y_* = R^{-1}y,\ X_* = R^{-1}X,\ \varepsilon_* = R^{-1}\varepsilon$$

$$y = X\beta + \varepsilon \text{ becomes}$$

$$R^{-1}y = R^{-1}X\beta + R^{-1}\varepsilon, \text{ or}$$

$$y_* = X_*\beta + \varepsilon_*$$

$$E(\varepsilon_*) = E(R^{-1}\varepsilon) = 0$$

Weighted least squares (continued)

$$Var(\varepsilon_*) = E\{[\varepsilon_* - E(\varepsilon_*)][\varepsilon_* - E(\varepsilon_*)]'\}$$

$$= E(\varepsilon_* \varepsilon_*')$$

$$= E(R^{-1}\varepsilon\varepsilon' R^{-1})$$

$$= R^{-1}E(\varepsilon\ \varepsilon')R^{-1}$$

$$= \sigma^2 R^{-1} V R^{-1}$$

$$= \sigma^2 R^{-1} R R R^{-1}$$

$$= \sigma^2 I$$

Weighted Least Squares (continued)

$$Q(\beta) = \varepsilon_*'\varepsilon_* = \varepsilon V^{-1}\varepsilon = \varepsilon W\varepsilon, \ \ W = V^{-1} = \textit{weights}$$

$$= (y - X\beta)'W(y - X\beta)$$

Least squares normal equations are $(X'WX)\hat{\beta} = X'Wy$

The solution is : $\hat{\beta} = (X'WX)^{-1}X'Wy$

$$Var(\hat{\beta}) = (X'WX)^{-1}X'W \,\mathrm{var}(y) WX(XWX)^{-1}$$

$$= \sigma^2(X'WX)^{-1}X'WW^{-1}WX(X'WX)^{-1}$$

$$= \sigma^2(X'WX)^{-1}$$

Robust Regression

Used to reduce influence of outliers

LAR Regression:

$$\text{minimize } \ L1 = \sum_{i=1}^{n} | y_i - x_i \beta | = \sum_{i=1}^{n} | e_i |$$

LMS Regression:

$$\text{minimize}: \text{median}\{[y_i - x_i \beta]^2\} = \text{median}\{e_i^2\}$$

M estimators:

$$\text{minimize}: \sum_{i=1}^{n} g(y_i - x_i \beta) = \sum_{i=1}^{n} g(e_i), \ \text{g a function of residuals}$$

Robust Regression (continued)

IRLS, iteratively reweighted least squares

Minimize e'We

W is a diagonal matrix of weights, inversely proportional to magnitude of scaled residuals, u_i

$u_i = e_i/s$, $s = MAD = median\{|e_i - median(e_i)|\}$

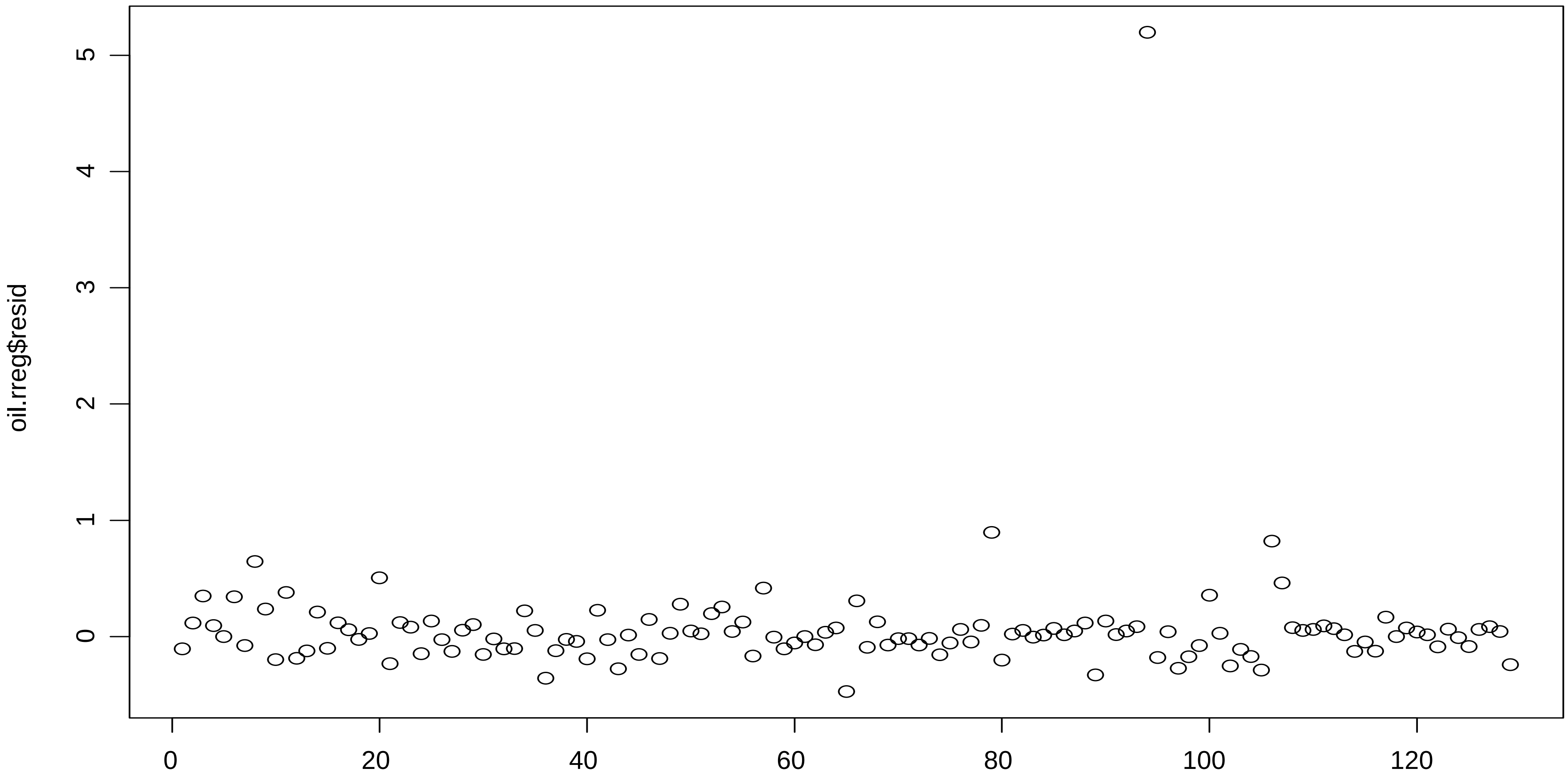
Procedure:

1. Obtain initial coefficient estimates from OLS
2. Obtain weights from scaled residuals
3. Obtain coefficient estimates from WLS
4. Return to 2.

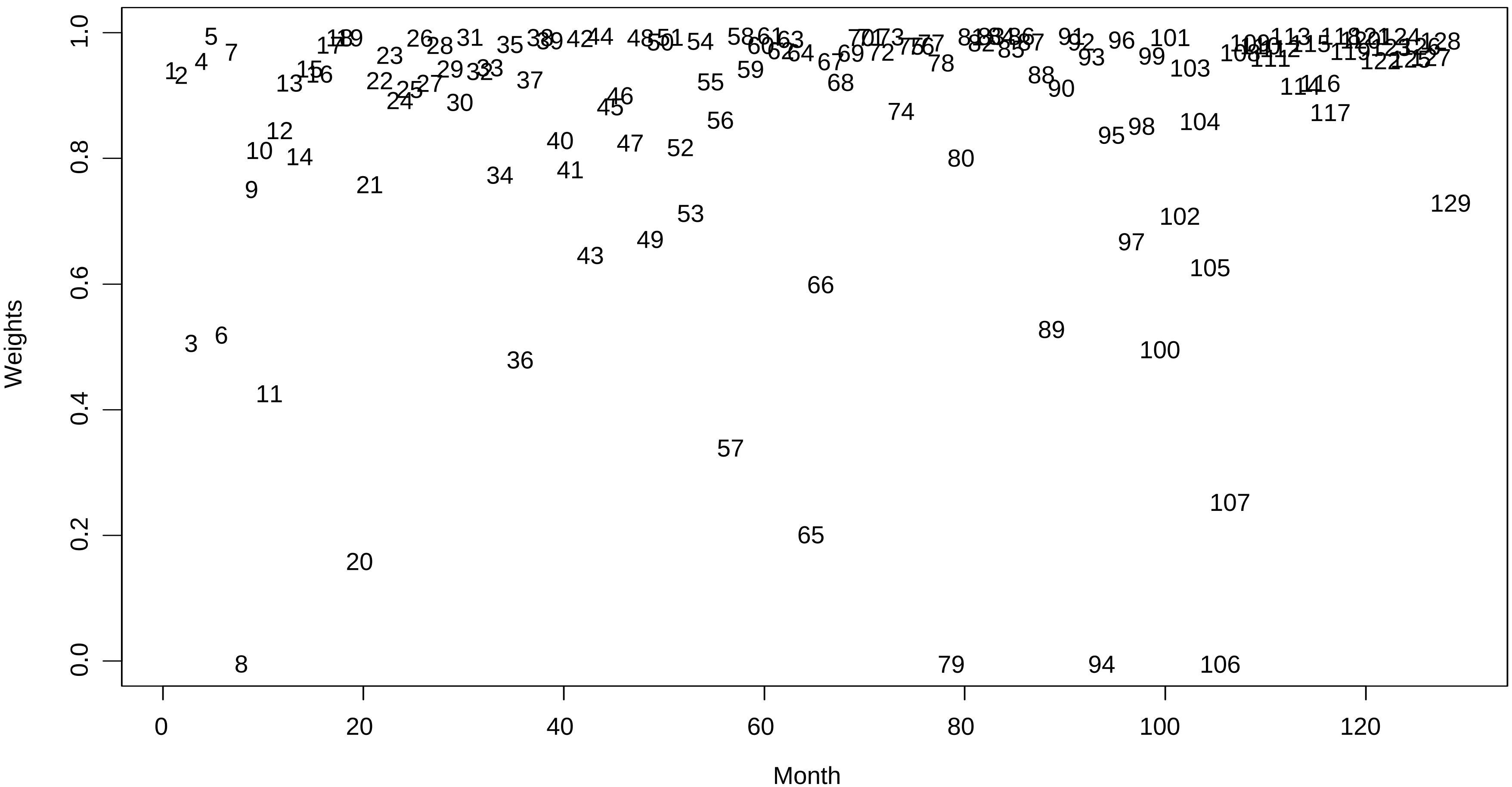
Convergence usually rapid.

(See Figure 10.4, and Equations 10.44 and 10.45 in Neter et al. *Applied Linear Statistical Models*.)

Plot of residuals in oil.rreg

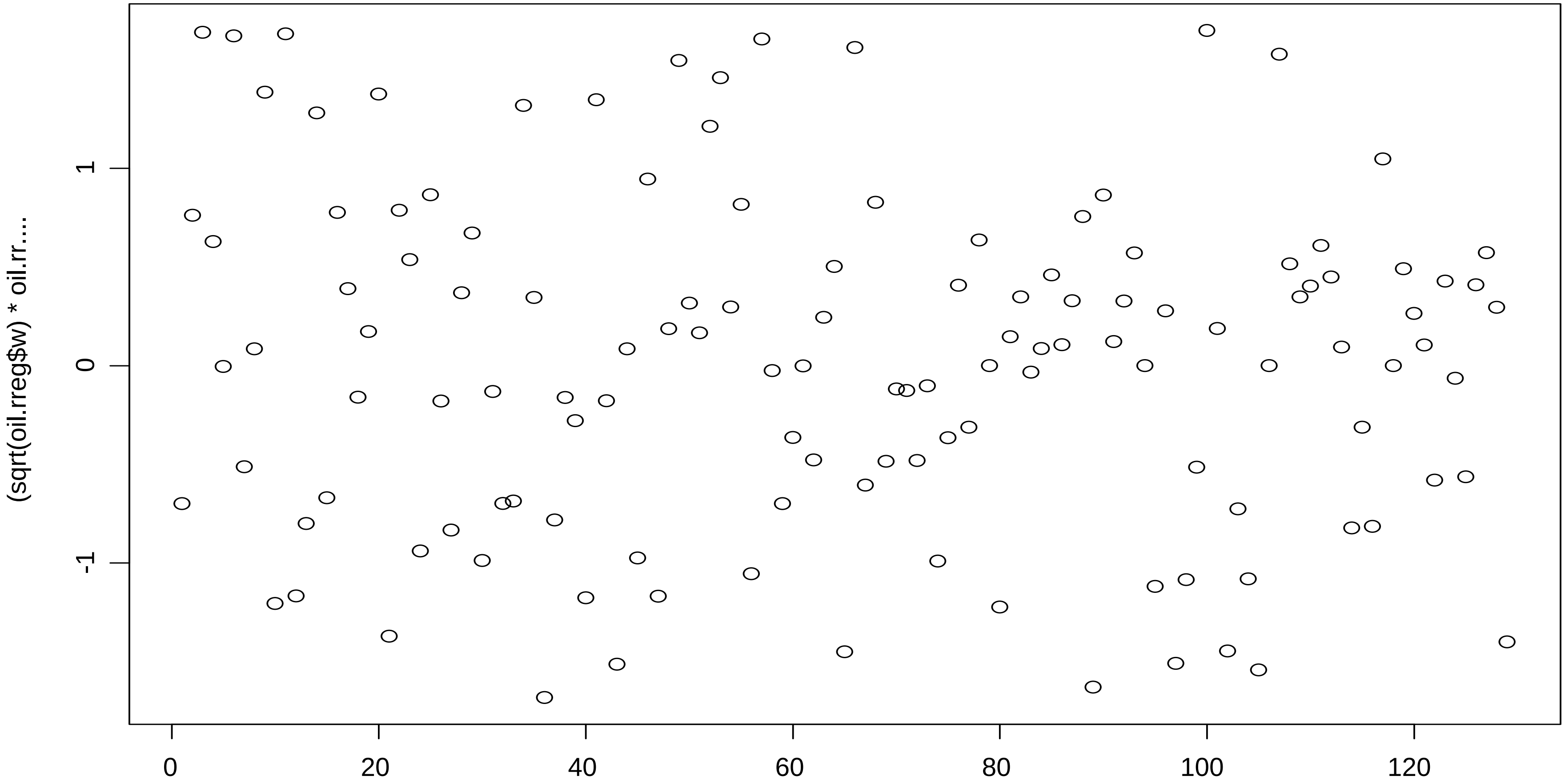


Plot of weights in robust regression for oil city data set



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

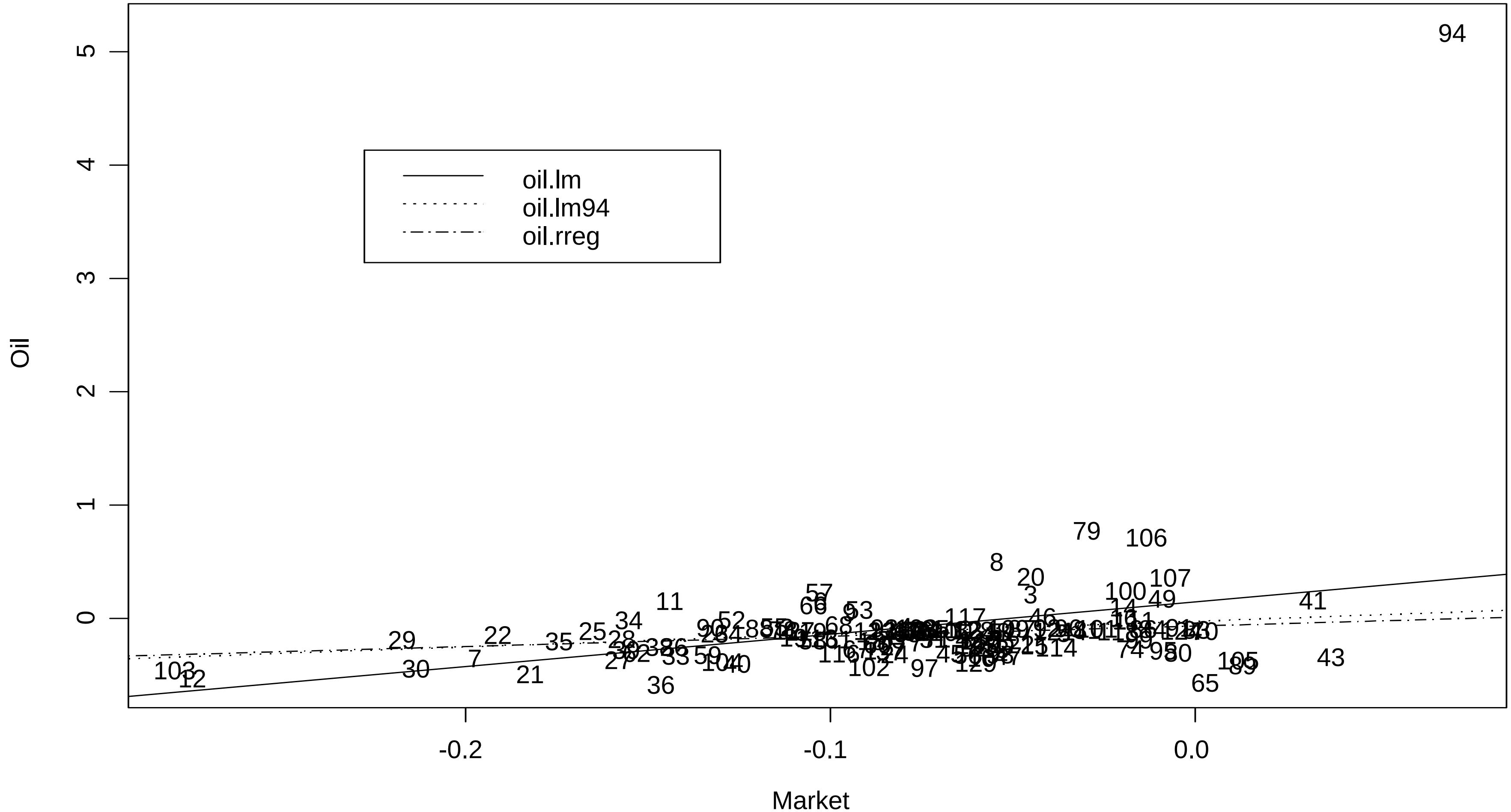
Plot of sqrt(weights)*resid/s in oil.rreg



Coefficient table for oil.rreg

```
> x<-cbind(1,Market)
> beta<-solve(t(x)%*%diag(w)%*%x)%*%t(x)%*%diag(w)%*%Oil
> r<-Oil-x%*%beta
> s<- median(abs(r-median(r)))*1.4826
> covm<-solve(t(x)%*%diag(w)%*%x)*s^2
> se<-sqrt(diag(covm))
> tvalue=beta/se
> prob<-2*(1-pt(abs(tvalue),127))
> cbind(beta,se,tvalue,prob)
                  beta         se    tvalue         prob
(Intercept) -0.06779903 0.02451469 -2.765649 0.0065285939
          x  0.89895511 0.24902845  3.609849 0.0004394276

Covariance matrix is approximate.
```

Plots of fitted regression lines for oil city data



Least Trimmed Squares Regression

Minimizes : $\sum_{i=1}^{q} e_i^2$,

where q is chosen to be between n/2 and n

Based on a genetic algorithm for finding a subset of data with minimum SSE.

High breakdown point: fits the bulk of the data well, even if bulk is only a little more than half the data.

Resulting weights are 1 or 0

```
> summary(oil.lts)
Method:
[1] "Least Trimmed Squares Robust Regression."

Call:
ltsreg(formula = Oil ~ Market)

Coefficients:
 Intercept  Market
 -0.0864    0.7907

Scale estimate of residuals: 0.1468

Robust Multiple R-Squared: 0.09863

Total number of observations:  129

Number of observations that determine the LTS estimate:  116

Residuals:
   Min. 1st Qu. Median 3rd Qu.   Max.
 -0.454 -0.088   0.032  0.097   5.223

Weights:
  0   1
 10 119
```