Regression

Regression is the most widely used of all statistical model. For univariate regression, the available data are one response variable and p predictor or explanatory variables. Let y_t and $x_{t1}, \dots, x_{tp}$, $t = 1, \dots, n$ denote a sample of n observations. Regression is the search for the functional relationship of the form $y \approx \varphi(\boldsymbol{x})$. The main goals of regression modeling include investigating how Y is related to $x_1, \dots, x_p$, estimating the conditional expectation of Y given $x_1, \dots, x_p$ and predicting future y value when the corresponding $x_1, \dots, x_p$ are available.

Linear Regression Models

The most popular regression model is the linear regression model in which the dependence function $\varphi(\cdot)$ is a linear function and the model,

$$y_t = \beta_0 + \beta_1 x_{t,1} + \cdots + \beta_p x_{t,p} + \varepsilon_t$$

where $\beta_0, \dots, \beta_p$ are fixed but unknown parameters, the Y_t and usually (but not always) x_{tk} are assumed to be random variables. The ε_t are often referred to as errors, it is assumed that

$$E(\varepsilon_t | x_{t,1}, \dots, x_{t,p}) = 0$$

so that

$$E[y_t \,|\, x_{t,1}, \dots, x_{t,p}] = \beta_0 + \beta_1 x_{t,1} + \cdots + \beta_p x_{t,p}\,.$$

It is more convenient to write our data in vectors and matrix as

$$y_t = \boldsymbol{x}_t \boldsymbol{\beta} + \varepsilon_t \,, \qquad , t = 1, \ldots, n,$$

where $\boldsymbol{x}_t = (x_{t,1}, \ldots, x_{t,p})^T$. and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$. The usual practice is to write down all the observations in one equation. Let

$$\underset{n\times 1}{\boldsymbol{y}} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} , \qquad \underset{n\times(p+1)}{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} , \qquad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} ,$$

the complete sample of n equations is represented by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \,. \tag{1}$$

Least Squares Estimator The method of least squares is the standard technique for extracting an estimate of β from a sample of observations. Consider

$$\underset{n\times 1}{\boldsymbol{e}(\boldsymbol{\beta})} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \,,$$

a vector of prediction error functions with $p+1$ dimensional vector argument $\boldsymbol{b}$ and the squared prediction errors is

$$S(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{e}(\boldsymbol{\beta})^T\boldsymbol{e}(\boldsymbol{\beta}).$$

The least squares estimator of $\boldsymbol{\beta}$ is the vector $\hat{\boldsymbol{\beta}}$ such that $S(\hat{\boldsymbol{\beta}}) \leq S(\boldsymbol{\beta})$ for all $p+1$ dimensional vector $\boldsymbol{\beta}$, that is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) \tag{2}$$

This is called the ordinary least squares (OLS) estimator and the solution of (2) is given by

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \,. \tag{3}$$

Throughout this handout, $\hat{\boldsymbol{\beta}}$ denotes $\hat{\boldsymbol{\beta}}_{\text{OLD}}$ unless it is stated otherwise. The least squares predictor of $\boldsymbol{y}$ or fitted value,

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{P}_X\boldsymbol{y} \tag{4}$$

the notation $\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the projection matrix, it is also called *hat matrix* in statistics. The residuals,

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y} \,. \tag{5}$$

It is straight forward to check that $\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{y}} = 0$ which implies the sum of squares decomposition,

$$\boldsymbol{y}^T\boldsymbol{y} = \hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} \,.$$

If we further assume that $\varepsilon_1, \ldots \varepsilon_n$ are $i.i.d. \sim (0, \sigma_\varepsilon^2)$, that is,

$$E[\boldsymbol{\varepsilon}\,|\boldsymbol{X}] = \boldsymbol{0} \,, \qquad E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\,|\boldsymbol{X}] = \sigma_\varepsilon^2\mathbf{I}_n \,. \tag{6}$$

Then the unbiased estimator of σ_ε^2 is,

$$\hat{\sigma}_\varepsilon^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{n-p-1} = \frac{1}{n-p-1}\sum_{t=1}^{n}(y_t - \boldsymbol{x}_t^T\hat{\boldsymbol{\beta}})^2 \,. \tag{7}$$

The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\text{Cov}(\hat{\beta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \sigma_\varepsilon^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \,.$$

Goodness of Fit The square of the sample coefficient between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ has the formula

$$R^2 = \frac{(\boldsymbol{y}^T\hat{\boldsymbol{y}} - n\bar{y}^2)^2}{(\boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2)(\hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}} - n\bar{y}^2)} = \frac{\hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}} - n\bar{\boldsymbol{y}}^2}{\boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2} = 1 - \frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{\boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2}\ .$$

This is known as *the coefficient of determination* and is conventionally used to measure the goodness of fit of the regression. A quantity related to R^2, known as *adjusted R squared* is also used for measuring goodness of fit. It is R^2 with the two sum of squares adjusted for their degrees of freedom,

$$R^2_{Adj} = 1 - \frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}/(n-p-1)}{(\boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2)/(n-1)} = 1 - \frac{\hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_y}$$

MLE for Regression If ε_t are $i.i.d.(0,\sigma^2_\varepsilon)$ with density $f(\cdot)$, then the likelihood of $\boldsymbol{y} = (y_1,\ldots,y_n)$ is

$$\prod_{t=1}^n \frac{1}{\sigma_\varepsilon} f\left(\frac{y_t - \boldsymbol{x}_t^T\boldsymbol{\beta}}{\sigma_\varepsilon}\right).$$

The maximum likelihood estimator maximizes the log-likelihood

$$L(\boldsymbol{\beta},\sigma_\varepsilon) = -n\log\sigma_\varepsilon + \sum_{t=1}^n \log f\left(\frac{y_t - \boldsymbol{x}_t^T\boldsymbol{\beta}}{\sigma_\varepsilon}\right)$$

For normally distributed errors, ignoring the constant term, the log likelihood is

$$L^{\text{NORM}}(\boldsymbol{\beta},\sigma_\varepsilon) = -n\log\sigma_\varepsilon - \frac{1}{2}\sum_{t=1}^n\left(\frac{y_t - \boldsymbol{x}_t^T\boldsymbol{\beta}}{\sigma_\varepsilon}\right)^2.$$

It should be obvious that the least-squares estimator is the MLE of $\boldsymbol{\beta}$ in the Normal case. The difference is the estimator of σ^2_ε,

$$\hat{\sigma}^2_{\varepsilon,MLE} = \frac{1}{n}\sum_{t=1}^n (y_t - \boldsymbol{x}_t^T\hat{\boldsymbol{\beta}})^2, \tag{8}$$

which is biased, we can always replace (4) with the unbiased estimator in (7).

Generalized Least Squares Estimator Often $\boldsymbol{\varepsilon}$ has a covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ other than identity for some density function $f(\cdot)$

$$|\Sigma_\varepsilon|^{-1/2} f\left\{(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

Then the likelihood is

$$-\frac{1}{2}\log|\Sigma_\varepsilon| + \log f\left\{(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}_\varepsilon^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

In the important special case where $\boldsymbol{\varepsilon}$ has a mean-zero multivariate normal distribution, the density of $\boldsymbol{\varepsilon}$ is

$$\frac{1}{|\Sigma_\varepsilon|^{1/2}(2\pi)^{p/2}}\exp\left\{-\frac{1}{2}\boldsymbol{\varepsilon}^T\Sigma_\varepsilon^{-1}\boldsymbol{\varepsilon}\right\}$$

If Σ_ε is known, then the MLE of $\boldsymbol{\beta}$ minimizes

$$(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}_\varepsilon^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$$

and is called the generalized least-squares estimator (GLS estimator) and the solution is given by

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\boldsymbol{X}^T\Sigma_\varepsilon^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\Sigma_\varepsilon^{-1}\boldsymbol{y}$$

Typically, Σ_ε is unknown and must be replaced by an estimate, for example, from an ARMA model for the errors.

Tests in the Linear Regression Models

Distribution of $\hat{\boldsymbol{\beta}}$ Consider the regression model (1) with assumption (6). Furthermore, assume additionally that

$$E\|\boldsymbol{x}_t \varepsilon_t\|^{2+\delta} < \infty\ , \qquad \delta > 0, \qquad \forall t,$$

then $\hat{\beta} = \hat{\beta}_{\text{OLS}}$ of (3) is normally distributed with the expected value $\boldsymbol{\beta}$ and variance $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ stated earlier.

t-test Let $\boldsymbol{M} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and m_{ii} the ith diagonal entry of $\boldsymbol{M}$, the t-statistics for testing $H_0 : \beta_i = b_i$ is

$$t = \frac{\hat{\beta}_i - b_i}{\hat{\sigma}_\varepsilon \sqrt{m_{ii}}} \sim t_{n-p-1} \qquad \text{under} \quad H_0\,.$$

The standard regression output gives the tests of $H_0 : \beta_i = 0$.

F-test Let $\boldsymbol{C}$ be a full rank constant matrix of dimension $r \times p$ and $\boldsymbol{c}$ a $r \times 1$ constant vector. The general form of the null hypothesis is

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{c}$$

The F statistic on H_0 is given by

$$\frac{1}{r\hat{\sigma}^2_\varepsilon}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T\{\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T\}^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \sim F_{r,n-p}\ .$$

The test result of $\boldsymbol{C} = \mathbf{I}_p$ and $\boldsymbol{c} = \boldsymbol{0}$ is a standard output of regression.

Eg 0.1. The manufacturing economy. The data consist of three Industrial Production (IP) indexes, and Capacity Utilization (CU) from Jan 2004 to Aug 2024. The IP index gives actual U.S. monthly manufacturing output, the second part gives percentage of production infrastructure that is being used; also a measure of potential output.

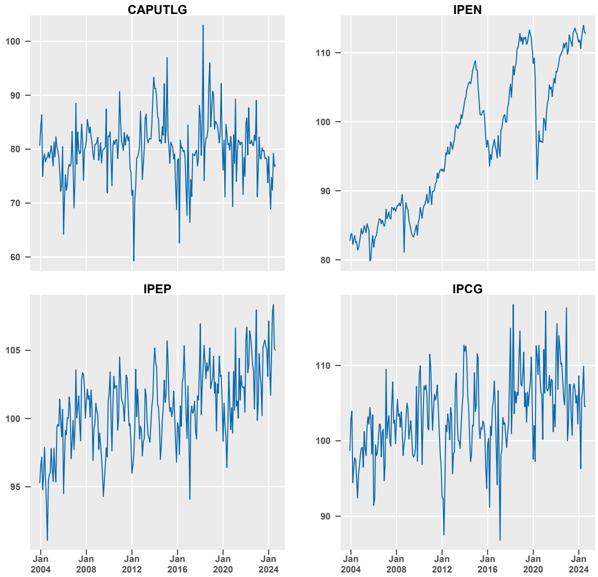

Figure 1: Time series plots of 4 macro economic series, from Dec 2003 to Aug 2024.

We will use the CP of utilities as response variable and the three IP indexes, electric power, total energy and consumer goods as regressors. When dealing with time series data such as macroeconomic series, we should examine the data with care before fitting.

Spurious Regression Granger and Newbold (1974) conducted a simulation study on independent random walks generated by independent Gaussian noise. A simulation study of the regression of one random walk on the other has a rejection rate of 76% on testing the null hypothesis that the slope $\beta_1 = 0$ using conventional t-statistics. Phillips (1986) showed that $\hat{\beta}_0$ and $\hat{\beta}_1$ do not converge in probability to constants as $n \to \infty$ and the conventional t-ratios do not have limiting distributions, in fact diverge as$n \to \infty$, so that there are no asymptotically correct critical values for these tests.

Cointegration Two integrated time series may share common stochastic trend so that the linear combination of them becomes stationary series. Such relationship is called cointegration. To determine the existence of cointegration requires several tests. If there is cointegration between the two series, the error correction model can be used for the bi-or-multi variate series.

These topic will be discuss later. The necessary tests have been given on the 4 series, we only have one nonstationary series, IPEN, that requires differencing. Both IPEP and IPCG have a linear time trend and CAPUTLG is stationary. We will remove the linear trend from IPEG and IPCG, and difference the IPEN series.

```
head(Wt,2); n0 = dim(Wt)[1]

##            CAPUTLG    IPEP    IPEN     IPCG
## 2003-12-01 80.6970 95.2977 82.7964  98.7142
## 2004-01-01 84.1209 96.5459 83.7058 102.9149


IPEP = lsfit(1:n0, Wt[,"IPEP"])$res  ## remove the linear trend
IPCG =  lsfit(1:n0, Wt[,"IPEP"])$res ## remove the linear trend
IPEN = diff(Wt[,"IPEN"]) ## difference the series
```

The function `lsfit()` is a simple version or R function to fit linear model, we use it because we only need the residuals from the least squares fit. Differencing a series will lose the first observation, the new data will start from Jan 2004.

```
yt = cbind(Wt[-1,"CAPUTLG"], IPEP[-1], IPEN, IPCG[-1])
colnames(yt) = syb
head(yt,2)

##            CAPUTLG       IPEP   IPEN       IPCG
## 2004-01-01 84.1209 -1.2409794 0.9094 -1.2409794
## 2004-02-01 86.3516 -0.6284786 0.0871 -0.6284786

tail(yt,2)

##            CAPUTLG     IPEP    IPEN     IPCG
## 2024-07-01 76.7524 1.258922 -1.0320 1.258922
## 2024-08-01 77.0490 1.129023 -0.1519 1.129023

n = dim(yt)[1]; cat("sample size:", n)

## sample size: 248
```

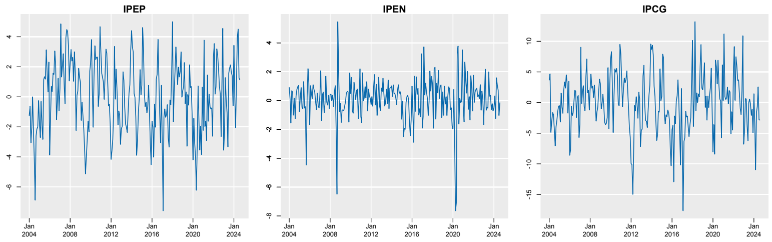

Figure 2: Time series plots of 3 IP indexes after removing nonstationarity, from Jan 2004 to Aug 2024.

The scatter plots of `yt` are also plotted. Focusing on the correlations at the first row, we find CP and IPCG has highest correlation.

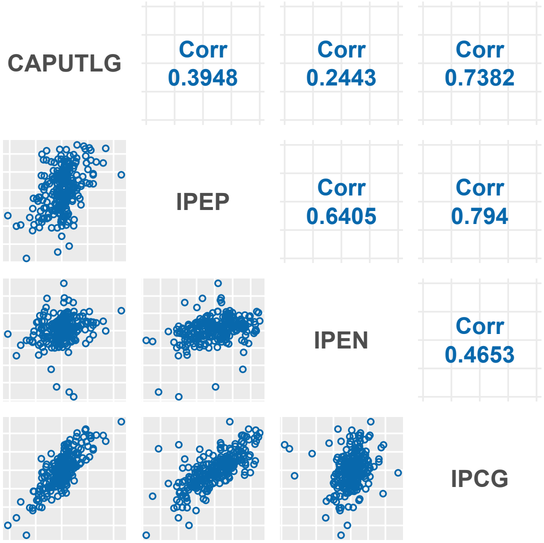

Figure 3: Scatter plots of the 4 series.

The regression analysis is done by R's `lm()` function.

```
args(lm)
## function (formula, data, subset, weights, na.action, method = "qr",
##     model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE,
##     contrasts = NULL, offset, ...)
y.dat = as.data.frame(yt)
lm.out = lm(CAPUTLG~IPEP + IPEN + IPCG, data = y.dat)
lm.out
##
## Call:
## lm(formula = CAPUTLG ~ IPEP + IPEN + IPCG, data = y.dat)
##
## Coefficients:
## (Intercept)         IPEP         IPEN         IPCG
##     80.1022      -0.8179       0.4735       1.2007
names(lm.out)
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

The return value includes fitted value $\hat{y}$ and residuals $\hat{\varepsilon}$. The tests can be obtained by `summary()` function.

```
summary(lm.out)
##
## Call:
## lm(formula = CAPUTLG ~ IPEP + IPEN + IPCG, data = y.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.491  -1.797  -0.027   1.955  11.184
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 80.10216    0.19994 400.639  < 2e-16 ***
## IPEP        -0.81792    0.12871  -6.355 1.02e-09 ***
## IPEN         0.47352    0.14460   3.275  0.00121 **
## IPCG         1.20070    0.06252  19.204  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.137 on 244 degrees of freedom
## Multiple R-squared:  0.6817,Adjusted R-squared:  0.6778
## F-statistic: 174.2 on 3 and 244 DF,  p-value: < 2.2e-16
lm.sum = summary(lm.out); names(lm.sum)

##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

The names of output value are self-explained, many of them are shown in the display.

Diagnostic Analysis We shall check assumptions and finding possible problems of the regression models. The function `plot()` will plot all six Diagnostic plots. The functions `cooks.distance()` and `hatvalues()` give Cook's statistics and leverage values.

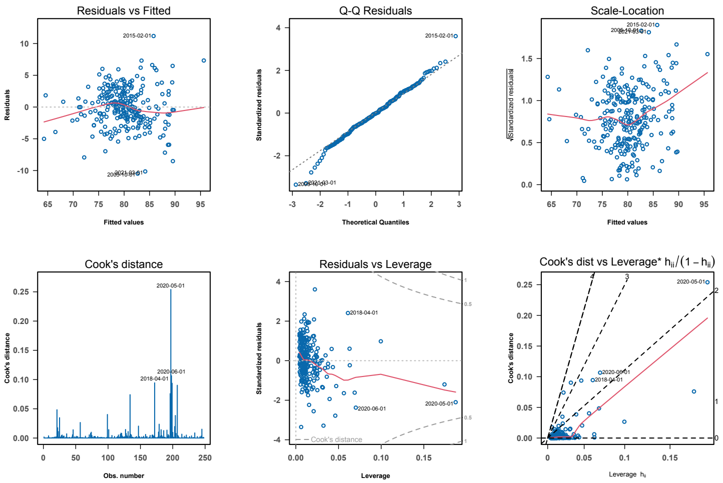

Figure 4: Diagnostic plot.

Leverages Let h_{tt} be the tth element of the diagonal entry of the hat matrix or projection matrix $\mathbf{P}_X$. The leverage of ith observation is h_{tt}, which depends only on $\boldsymbol{X}$, it is between 0 and 1 and have an average value of p/n. Points with h_{ii} greater than $2p/n$ are generally regarded as high leverage points.

Residuals We have defined the fitted values and residuals from fitting in (4) and (5) from fitting a least squares regression. The residuals defined in (5) are raw residuals, which have higher variance with higher leverage. The standard error of $\hat{\varepsilon}_t$ is $\hat{\sigma}_\varepsilon\sqrt{1-h_{tt}}$, which has higher variance with higher leverage. The standardized residuals are defined as

$$\hat{\varepsilon}_t^{(s)} = \frac{\hat{\varepsilon}_t}{\hat{\sigma}_\varepsilon\sqrt{1-h_{tt}}}\,.$$

The standardized residuals can help us identify outliers.

Cook's Distance An influence measure proposed by Cook (1977) is widely used. Cook's distance or Cook's D for the tth observation is given by

$$D_t = \frac{h_{tt}}{1-h_{tt}} \times \frac{{\hat{\varepsilon}_t^{(s)}}^2}{p}$$

The distance can be interpreted in several ways. One intuitive way is to think of it as the average squared difference between the predictions from the full and reduced data by deleting one observation compared. Algebraically, it can be shown that D_t defined above is the same as

$$D_t = \frac{\sum_{u=1}^{n}(\hat{y}_t - \hat{y}_{t(-u)})^2}{p\hat{\sigma}_\varepsilon^2}\,,$$

where $\hat{y}_{t(-u)}$) is the fitted values from the least squares fit without the uth observation. A large value of D_t indicates that the point is influential. However, there is no test for determining influence. There were suggestions of cut offs (eg. 1) but do not work well. A more effective way is to examine the plot like the one given in Figure 4 .