Clevere Analysen in Oracle ohne kostenpflichtige Zusatzoptionen
DOAG 2015

DB-Systel GmbH

Dr. Bernd Günther

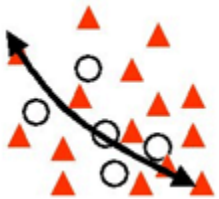
I.LPA46

Nürnberg, 18.11.2015

Wissensgewinnung aus Datenbanken

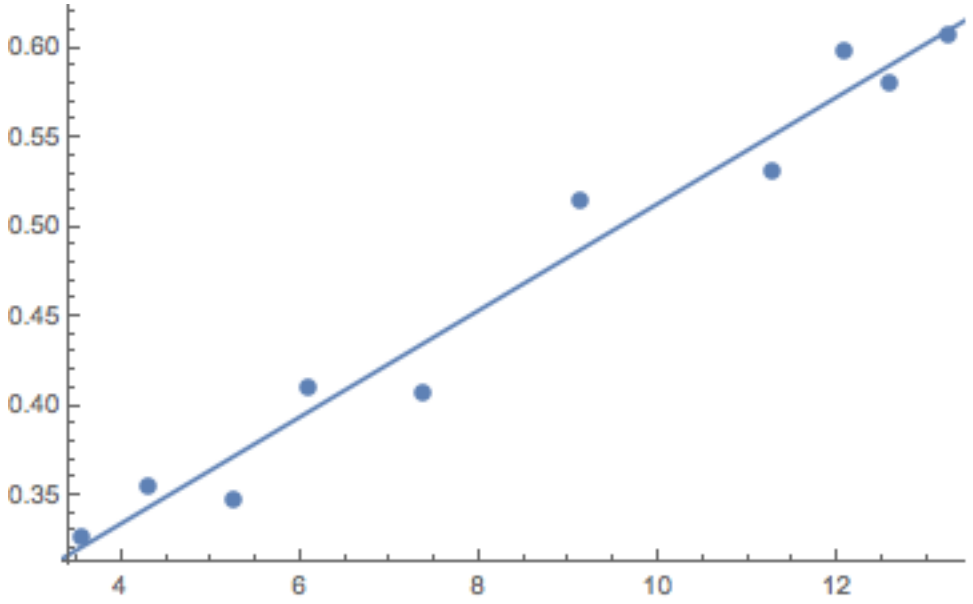
- Oracle Technology Network > Database > Options > Advanced Analytics

- Techniques:
  - **Regression**
  - Classification
  - Attribute Importance
  - Anomaly Detection
  - Clustering
  - Association
  - Feature Selection and Extraction



Anwendungsfall: von Schnittstellenpartner verursachte CPU-Last auf einem Datenbankserver

Abfragen/min x	CPU-Last y
11,27	50,96%
0,48	14,89%
9,16	46,75%
13,93	61,98%
10,42	54,39%
10,85	56,37%
5,53	36,82%
4,04	22,15%
4,50	32,90%
11,63	57,54%



Lösung mit SQL-Mitteln:

```
CREATE TABLE TAB4ANALYSIS1
(X NUMBER NOT NULL,
Y NUMBER NOT NULL);


SQL> select REGR_INTERCEPT (y, x), REGR_SLOPE (y, x) from tab4analysis1;

REGR_INTERCEPT(Y,X) REGR_SLOPE(Y,X)
------------------- ---------------
         .133958804     .036762477
```

Mögliche Erweiterungen:

- Mehr Schnittstellenpartner (unabhängige Variablen):

$$y = a_0 + a_1 x_1 + \cdots + a_n x_n$$

- Allgemeinere Funktionen (z.B. Polynome):

$$y = a_0 + a_1 x + \cdots a_n x^n$$

(das ist eine lineare Regression mit $x_1 = x$, $\ldots x_n = x^n$)

Mit einfachen SQL-Mitteln nicht mehr lösbar.

x_1	x_2	y
9,12	4,31	68,79%
4,30	3,31	48,79%
6,08	0,86	44,48%
7,37	7,42	70,42%
5,26	0,27	35,83%
13,23	9,87	90,26%
12,08	6,44	85,60%
3,55	8,92	68,36%
11,26	6,08	77,49%
12,58	9,85	97,49%

Methode 1: Oracle Datamining

```
CREATE TABLE TAB4ANALYSIS2(
X1 NUMBER NOT NULL,
X2 NUMBER NOT NULL,
Y NUMBER NOT NULL,
ID INTEGER PRIMARY KEY);

CREATE TABLE REGR_2DIM_SETTINGS
(SETTING_NAME VARCHAR2(30),
SETTING_VALUE VARCHAR2(4000));
```

Methode 1: Oracle Datamining

```
begin
insert into REGR_2DIM_SETTINGS values(dbms_data_mining.algo_name,
dbms_data_mining.algo_generalized_linear_model);
commit;

SYS.DBMS_DATA_MINING.CREATE_MODEL(
model_name => 'REGRESSION_2DIM_MODEL',
mining_function => dbms_data_mining.regression,
data_table_name => 'TAB4ANALYSIS2',
target_column_name => 'Y',
settings_table_name => 'REGR_2DIM_SETTINGS',
case_id_column_name => 'ID'
);
end;
/
```

Methode 1: Oracle Datamining

```
SQL> SELECT ATTRIBUTE_NAME, COEFFICIENT FROM
TABLE(DBMS_DATA_MINING.GET_MODEL_DETAILS_GLM('REGRESSION_2DIM_MODEL'));

ATTRIBUTE_NAME      COEFFICIENT
-------------------- -----------
                      .132174284
X1                    .036774777
X2                    .040349144
```

- Literatur: Brendan Tierney, Predictive Analytics Using Oracle Data Miner: Develop & Use Data Mining Models in ODM, SQL & PL/SQL
- **Nachteil dieser Methode: Erwerb der Oracle Advanced Analytics Option ist erforderlich.**

Methode 2: Regression im Eigenbau

- Literatur z.B.: Ansgar Steland, Basiswissen Statistik, 3. Aufl.

- Lineare Regression erfordert zwei Hilfsmittel:
  - Aggregationen (Oracle: Aggregatfunktionen SUM, AVG etc.)
  - Lösen linearer Gleichungen (Oracle: Package UTL_NLA)
    - http://www.netlib.org/blas/ (Basic linear algebra subprograms)
    - http://www.netlib.org/lapack/ (Linear algebra package)

Methode 2: Regression im Eigenbau
Schritt 1

```
SELECT
AVG(X1*X1),
AVG(X1*X2),
AVG(X2*X1),
AVG(X2*X2),
AVG(X1),
AVG(X2),
AVG(X1*Y),
AVG(X2*Y),
AVG(Y)
FROM TAB4ANALYSIS2;
```

Methode 2: Regression im Eigenbau
Schritt 2

Ermitteln Sie die Koeffizienten a_0, a_1, a_2 aus dem folgenden Gleichungssystem

```
a0*AVG(X1)  +  a1*AVG(X1*X1)  +  a2*AVG(X1*X2)  = AVG(X1*Y)

a0*AVG(X2)  +  a1*AVG(X2*X1)  +  a2*AVG(X2*X2)  = AVG(X2*Y)

a0  +          a1*AVG(X1)  +     a2*AVG(X2)  =    AVG(Y)
```

mit Hilfe der Funktion `UTL_NLA.LAPACK_GESV`.

Representativität des Datenbestandes
Anforderung 1

Gut:

$$\frac{\mathrm{STDDEV}(x)}{\mathrm{AVG}(x)} \gg 0$$

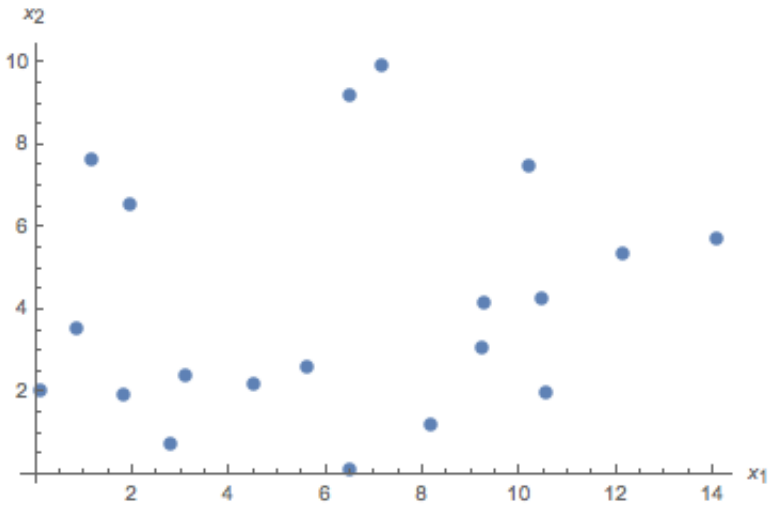
Schlecht:

$$\frac{\mathrm{STDDEV}(x)}{\mathrm{AVG}(x)} \sim 0$$

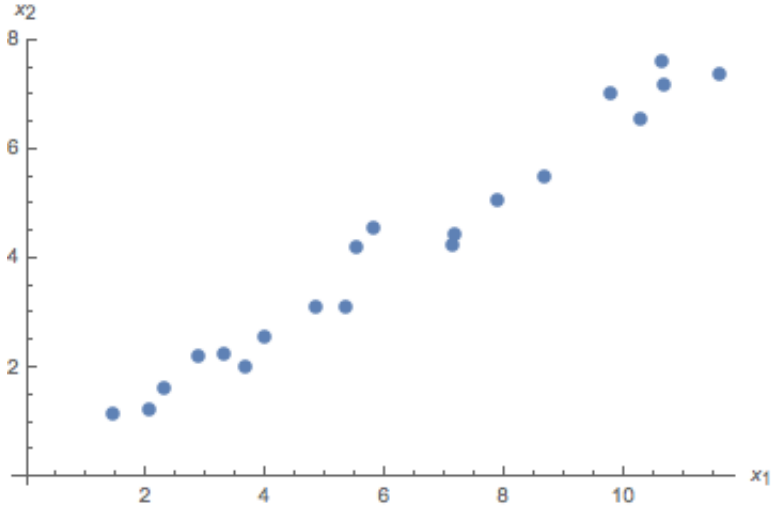
In mehreren Dimensionen muss diese Bedingung für jede einzelne *unabhängige* Koordinate $x_1, \cdots x_n$ erfüllt sein.

Representativität des Datenbestandes
Anforderung 2 (dim≥ 2)

Gut:

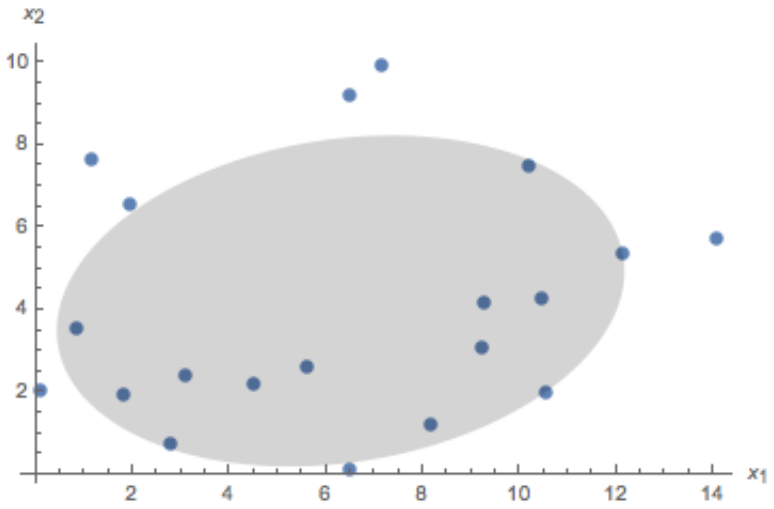


Schlecht:

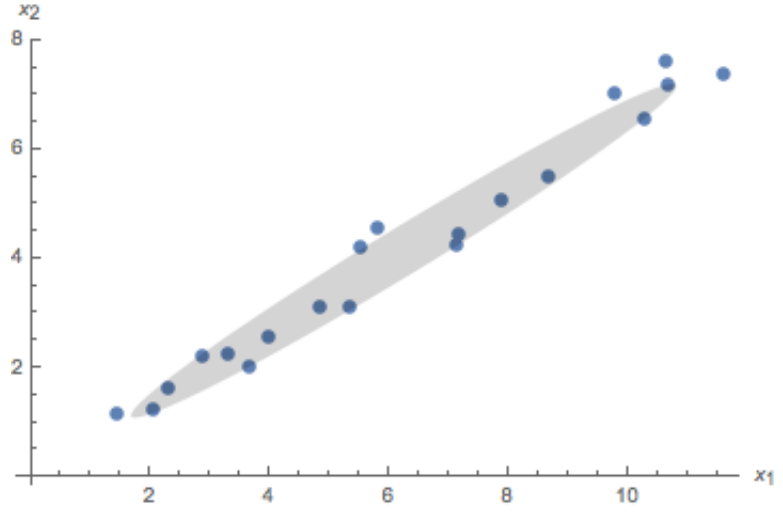


Representativität des Datenbestandes
Anforderung 2 (dim≥ 2)

Gut:



Schlecht:



Schwerpunkt in den Ursprung verschieben, Standardabweichungen auf Eins normieren; dann definiert die Koeffizientenmatrix $\begin{pmatrix} \mathrm{avg}(x_1x_1) & \mathrm{avg}(x_1x_2) \\ \mathrm{avg}(x_2x_1) & \mathrm{avg}(x_2x_2) \end{pmatrix}$ von Seite 11 das *Trägheitsellipsoid* der Punktwolke.

Representativität des Datenbestandes
Anforderung 2 (dim≥ 2)

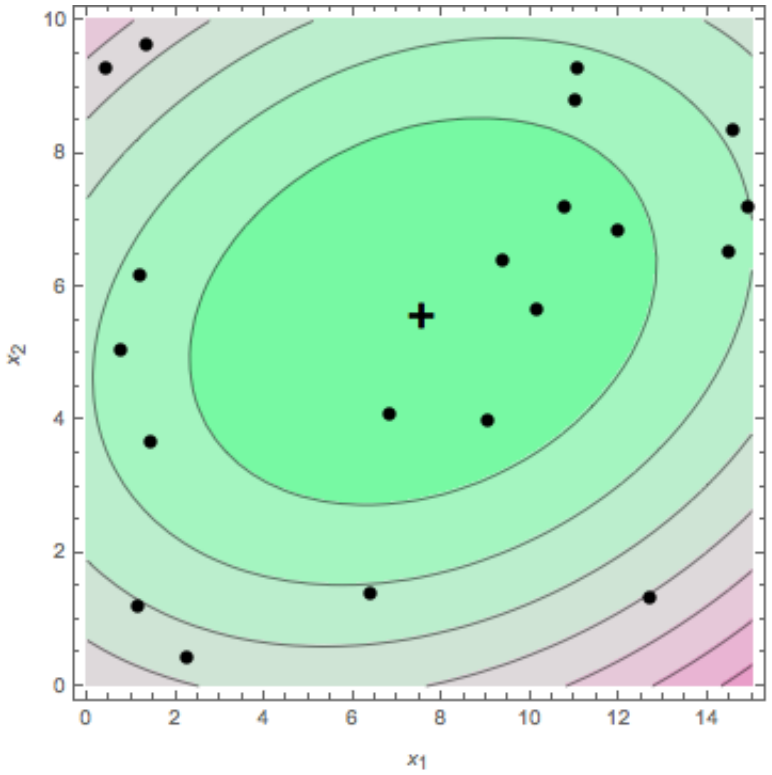
Die Koeffizientenmatrix der Prozedur

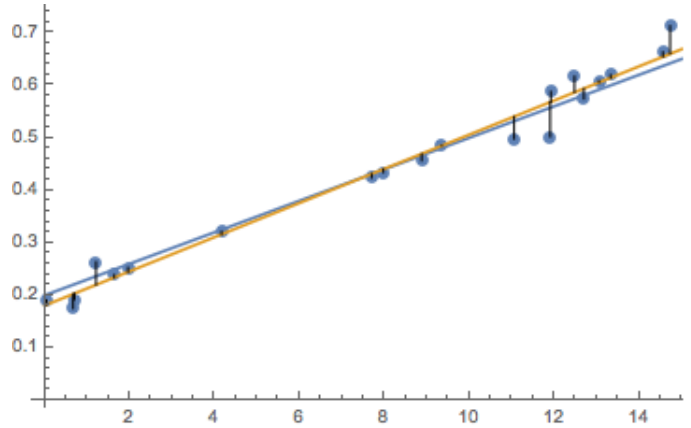
`UTL_NLA.LAPACK_SYEV`

übergeben. Als Rückgabe erhält man Länge und Lage der Halbachsen des Trägheitsellipsoids.

Unschärfe bei der Bestimmung der Regressionsgeraden

- Zwei Komponenten tragen zur Unschärfe einer Regressionsschätzung bei:
  - Streuung der Datenpunkte um die Regressionsgerade (bzw. -ebene, -hyperebene)
  - Unsicherheit über die Regression selbst (d.h. über die ermittelten Parameter)





Die Unschärfe ist im Datenschwerpunkt am kleinsten und nimmt mit der Entfernung zu. Mathematisch ergibt sich die Abschätzung aus der inversen Matrix der Koeffizientenmatrix von Seite 11.
→ Erneuter Aufruf der Funktion
`UTL_NLA.LAPACK_GESV`

Fragen?