STATISTISCHE ÄMTER
DER LÄNDER
FORSCHUNGSDATENZENTRUM

Zu den Enthüllungsrisiken der Regressionsanalyse beim Remote Access

14. Januar 2011, Wiesbaden

Dr. Alexander Vogel

Überblick

1. Motivation
2. Kurzexkurs: Geheimhaltung von Merkmalssummen
3. Grundidee
4. Enthüllung von Merkmalssummen nicht gezielt erstellter Variablen
5. Enthüllung mit gezielt erstellten (strategischen) Variablen
6. Fazit

Motivation I

- Mikrodatennutzung mittels kontrollierter Datenfernverarbeitung rückt immer stärker in den Fokus
- Zukunftsvision: vollautomatisiertes Fernrechnen
- Systematische Suche nach Enthüllungsrisiken durch multivariate Analysemethoden notwendig
- Im Rahmen des infinitE-Projektes: Ronning et al (2010) zeigen u.A.
  - Setzen von strategischen Dummies
  - Erzeugen künstlicher Ausreißer
  - …

Motivation II

- Ronning et al. (2010):

- „Die hier präsentierten Beispiele sind sicher nur ein kleiner Ausschnitt dessen sind, was möglich ist, vor allem wenn mathematisch-statistisch Begabte sich mit diesem Problem beschäftigen“

➢ Präsentation einer weiteren Möglichkeit, geheim zu haltende Werte mittels Regressionsoutput offenzulegen

Exkurs: Geheimhaltung von Merkmalssummen

- Primäre Geheimhaltung: Sperrung von Merkmalssummen durch
  - Mindestfallzahlregeln
  - Dominanzregeln
  - p%-Regel

- Zusätzliche Sperrungen von Merkmalssummen durch sekundäre Geheimhaltung

Grundidee

- Regressionsgerade verläuft durch den Schnittpunkt der Mittelwerte:

$$\bar{y} = b_0 + b_1\bar{x}_1 + b_2\bar{x}_2 \ldots + b_k\bar{x}_k + b_z\bar{z}$$

  - Abhängige Variable *y*,
  - *k* unabhängige Variablen x_1 bis x_k ,
  - geschätzten Regressionskoeffizienten b_0 bis b_k ,
  - *z* zeigt Variable deren Merkmalssumme für die in der Regression betrachtete Gruppe von Merkmalsträgern geheim gehalten werden muss

Grundidee

- Herleitung:

Diese herkömmliche Methode ist die Minimum-Quadrat-Methode oder Methode der kleinsten Quadrate. Man minimiert die summierten Quadrate der Residuen,

$$RSS = SS_{\text{Res}} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 \rightarrow \min!$$

bezüglich a und b. Durch partielles Differenzieren und Nullsetzen der Ableitungen erster Ordnung erhält man ein System von Normalgleichungen.

Die gesuchten *Regressionskoeffizienten* sind die Lösungen

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}}$$

und

$$a = \bar{y} - b\bar{x}$$

Grundidee

- Mittelwert von *z* ergibt sich durch einfaches Umstellen:

$$\bar{z} = \frac{b_0 + b_1\bar{x}_1 + b_2\bar{x}_2 \ldots + b_k\bar{x}_k - \bar{y}}{-b_z}$$

- Merkmalssumme von *z* ergibt sich durch

$$\sum_{i=1}^{n} z_i = \bar{z} \cdot n$$

  - *n*: Anzahl der Beobachtungen die in Regression einfließen

Grundidee

- Reznek (2003) bezogen auf die Tatsache, dass die Regressions-gerade durch den Schnittpunkt der Mittelwerte verläuft:

- „if researchers have a sample size that allows meaningful regression results, then usually the sample will present no disclosure risks“

Grundidee

- Folgende 3 Fälle zeigen, dass ein Blick auf die in die Regression einfließenden Fallzahlen nicht genügt:

(1) z nimmt für sehr viele Beobachtungen den Wert Null an, nur für ein oder zwei Merkmalsträger liegt ein Wert ungleich Null vor -> Sperrung der Merkmalssumme aufgrund der Mindestfallzahlregel

(2) z nimmt für viele Merkmalsträger nur einen sehr kleinen Wert und für einen oder zwei Merkmalsträger einen sehr großen Wert an -> Sperrung der Merkmalssumme aufgrund der Dominanzregel

(3) Die Merkmalssumme von z fungiert in einer Veröffentlichung der amtlichen Statistik als Sekundärsperrpartner und ist daher geheim zu halten.

STATISTISCHE ÄMTER
DER LÄNDER
FORSCHUNGSDATENZENTRUM

Enthüllung von Merkmalssummen nicht gezielt erstellter Variablen

- Beispiel: Für Hamburg 2003 ist der Wert der Zuckerrübenanbaufläche gesperrt

Landwirtschaftliche Betriebe mit Ackerland und deren Ackerfläche nach Fruchtarten - Erhebungsjahr - regionale Tiefe: Bundesländer

Allgemeine Agrarstrukturerhebung

Bundesländer Landwirtschaftliche Betriebe mit Ackerland Landwirtschaftlich genutzte Fläche			Einheit	Hackfrüchte	
				Kartoffeln	Zuckerrüben
2007					
DG	Deutschland	Landwirtschaftliche Betriebe mit Ackerland	Anzahl	54 930	37 885
		Landwirtschaftlich genutzte Fläche	ha	274 961	402 697
01	Schleswig-Holstein	Landwirtschaftliche Betriebe mit Ackerland	Anzahl	526	848
		Landwirtschaftlich genutzte Fläche	ha	5 949	10 981
02	Hamburg	Landwirtschaftliche Betriebe mit Ackerland	Anzahl	13	1
		Landwirtschaftlich genutzte Fläche	ha	15	.
03	Niedersachsen	Landwirtschaftliche Betriebe mit Ackerland	Anzahl	6 584	7 [illegible]
		Landwirtschaftlich genutzte Fläche	ha	120 231	100 667

Enthüllung von Merkmalssummen nicht gezielt erstellter Variablen

- Offenlegung durch folgendes Regressionsmodell möglich:

- y: Standarddeckungsbeitrag der landwirtschaftlichen Betriebe in Euro (EF70)
- x_1: landwirtschaftlich genutzte Fläche (EF258)
- x_2: Anzahl der Rinder (EF119)
- x_3: Dummy, Einzelunternehmen 1=ja, 0=nein (EF13_einzel)
- z: Zuckerrübenfläche (EF220)

Enthüllung von Merkmalssummen nicht gezielt erstellter Variablen

- 1. Schritt Regressionsanalyse:

```
. reg EF70 EF258 EF119 EF13_einzel EF220
      Source |       SS       df       MS              Number of obs =    1117
-------------+------------------------------           F(  4,  1112) =    2.08
       Model |  2.9479e+11     4  7.3699e+10           Prob > F      =  0.0807
    Residual |  3.9316e+13  1112  3.5356e+10           R-squared     =  0.0074
-------------+------------------------------           Adj R-squared =  0.0039
       Total |  3.9611e+13  1116  3.5494e+10           Root MSE      =  1.9e+05
------------------------------------------------------------------------------
        EF70 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       EF258 |   443.6301   320.7652     1.38   0.167     -185.743    1073.003
       EF119 |   -517.899   343.9492    -1.51   0.132    -1192.762    156.9637
 EF13_einzel |  -51088.58   21037.29    -2.47   0.014    -92385.05    -9792.11
       EF220 |          x          x    -0.28   0.780            x           x
       _cons |   143833.1   20594.82     6.98   0.000      103424    184242.2
------------------------------------------------------------------------------
```

Enthüllung von Merkmalssummen nicht gezielt erstellter Variablen

- 2. Schritt: Ausgabe der unverdächtigen Mittelwerte der Variablen EF70, EF258, EF124, und EF13_einzel für alle landwirtschaftlichen Betriebe in HH

➢ Zuckerrübenfläche könnte (mit einer geringen Abweichung von 0.019 %) berechnet werden

- Alternativer 2. Schritt: Nutzer oder interessierter Leser holt sich Angaben über EF70, EF258, EF124, und EF13_einzel aus den Statistischen Berichten

➢ Zuckerrübenfläche kann immer noch mit einer relativ geringen Abweichung von 0.103 % berechnet werden

Enthüllungen mit gezielt erstellten (strategischen) Variablen

- Enthüllung von Werten einer bestimmten Einheit
  - Offenlegung des unbekannten Werts *z* einer bestimmten Einheit *m*
  - Erstellung einer transformierten Variante der Variable *z,* welche nur für die Einheit *m* den Wert z_m enthält und für alle anderen Einheiten den Wert Null annimmt.
  - Folgende zwei Möglichkeiten:

Enthüllungen mit gezielt erstellten (strategischen) Variablen

- Enthüllung von Werten einer bestimmten Einheit (a)
  - Datenangreifer kennt ein Merkmal *w*, von dem er weiß, dass die Ausprägung w_m nur einmal im Datensatz vorkommt (die Einheit *m* also eindeutig beschreibt).
  - Erzeugt wird nun eine Variable *z_transformiert* die nur dann die Werte der Variable *z* enthält, wenn das Merkmal *w* die Ausprägung w_m annimmt.

$$z\_transformiert = \begin{cases} z & falls\ w = w_m \\ 0 & sonst \end{cases}$$

  - ➢ Beispiel Zuckerrübenbauer in Hamburg

Enthüllungen mit gezielt erstellten (strategischen) Variablen

- Enthüllung von Werten einer bestimmten Einheit (b)
  - Datenangreifer kennt die Intervalle, in denen sich die Merkmalsausprägungen der Variablen w_1 bis w_k der anzugreifenden Einheit befinden
  - Bedingung: Datenangreifer weiß, dass die Kombination genügt, um die anzugreifende Einheit eindeutig zu identifizieren.
  - Ist die Einheit *m* damit eindeutig identifiziert, lässt sich für jede dem Datenangreifer unbekannte Variable *z* eine transformierte Variable analog zu Möglichkeit (a) erstellen

Enthüllungen mit gezielt erstellten (strategischen) Variablen

- Enthüllung einer geheim zu haltenden Merkmalssumme
  - Regressionsanalyse gesperrter Merkmalssummen kleiner Untergruppen (Wirtschaftszweige / tiefe regionale Ebene) führen zu (auffälligerer) kleiner Anzahl von Beobachtungen
  - Transformation von *z,* so dass *z_transformiert* nur dann Werte enthält, wenn die Einheit zur gewünschten Untergruppe gehört.

$$z\_transformiert = \begin{cases} z & falls\ kreis = gewünschter\ kreis \\ 0 & sonst \end{cases}$$

  - Hamburger Zuckerrübenfläche ließe sich in Regression mit über 400.000 Fällen verstecken

Fazit – notwendige Prüfungen

- Hier vorgestellte Eigenschaft der Regressionsgerade birgt Enthüllungsrisiko nicht nur durch strategisch erstellte Variablen sondern auch durch unauffälligere „natürlich“ vorkommende Konstellationen
- Notwendige Prüfungen um Enthüllung zu verhindern:
  - Bilden von strategischen Variablen muss im Rahmen der Syntax-Kontrolle unterbunden werden
  - Prüfung metrischer Variablen in Regressionen auf Mindestfallzahl der Nicht-Null-Werte, Dominanz und Sekundärsperrung
  - Prüfung ausgeschlossener Untergruppen (analog zu Tabellen)

Fazit – laufender Prozess

- Ronning et al. (2010):

- „Die hier präsentierten Beispiele sind sicher nur ein kleiner Ausschnitt dessen sind, was möglich ist, vor allem wenn mathematisch-statistisch Begabte sich mit diesem Problem beschäftigen“

STATISTISCHE ÄMTER DER LÄNDER
FORSCHUNGSDATENZENTRUM

Vielen Dank!

Kontakt

STATISTISCHE ÄMTER DER LÄNDER - Forschungsdatenzentrum

Standorte Hamburg und Kiel
Dr. Alexander Vogel, Tel: 0431/6895-9113

fdz@statistik-nord.de