

Lösung zu Kapitel 9: Beispiel 3

In Florida war der Ausgang US-Präsidentschaftswahl im Jahr 2000 zwischen George W. Bush und Al Gore sehr knapp und umstritten. Unter anderem wurde in einem County ein Wahlzettel verwendet, mit dem leicht statt einer Stimme für Gore eine Stimme für Buchanan abgegeben werden konnte.

Im Datenfile (`florida2000.dat`) sind für die 67 Counties von Florida die Stimmen für alle Kandidaten aufgelistet. Wir untersuchen die Stimmen für Bush und Buchanan.

- Identifizieren Sie im Streudiagramm den Ausreißer unter diesen Daten!
- Gibt es einen Zusammenhang zwischen der Anzahl an Stimmen für Bush und Buchanan?
- Schätzen Sie (ohne den Ausreißer) die erwartete Stimmenzahl für Buchanan aus der Stimmenzahl für Bush! Wie viele Stimmen hätte man für Buchanan im Ausreißer-County erwartet?

Nach Einlesen der Daten gilt unser erstes Interesse den Variablen im Datensatz:

R

```
> fl2000 <- read.table("florida2000.dat", header = TRUE)
> names(fl2000)
```

```
[1] "County"      "Gore"         "Bush"         "Buchanan"     "Nader"
[6] "Brown"       "Hagelin"      "Harris"       "McReynolds"   "Moorehead"
[11] "Phillips"    "Total"
```

R

```
> attach(fl2000)
```

Offenbar sind die Stimmen für Bush in der Variablen `Bush`, jene für Buchanan in der Variablen `Buchanan` enthalten. Mit einem Streudiagramm machen wir uns ein Bild von den Daten, dabei verkleinern wir mit `cex=0.8` die Punktbeschriftung etwas:

R

```
> plot(Bush, Buchanan)
> text(Bush, Buchanan, County, cex = 0.8)
```

Es ist klar ersichtlich (► Abbildung 1), dass für die meisten Counties (auch wenn sich die Beschriftungen einander teilweise überdecken) die Relation zwischen der Stimmenanzahl für Bush und jener für Buchanan ähnlich ist. Das County Palm Beach weicht davon deutlich ab.

Wir merken uns, wieviel Stimmen Bush und Buchanan im County Palm Beach erhalten haben (`BushPB` und `BuchananPB`).

R

```
> BushPB <- Bush[County == "Palm Beach"]
> BuchananPB <- Buchanan[County == "Palm Beach"]
```

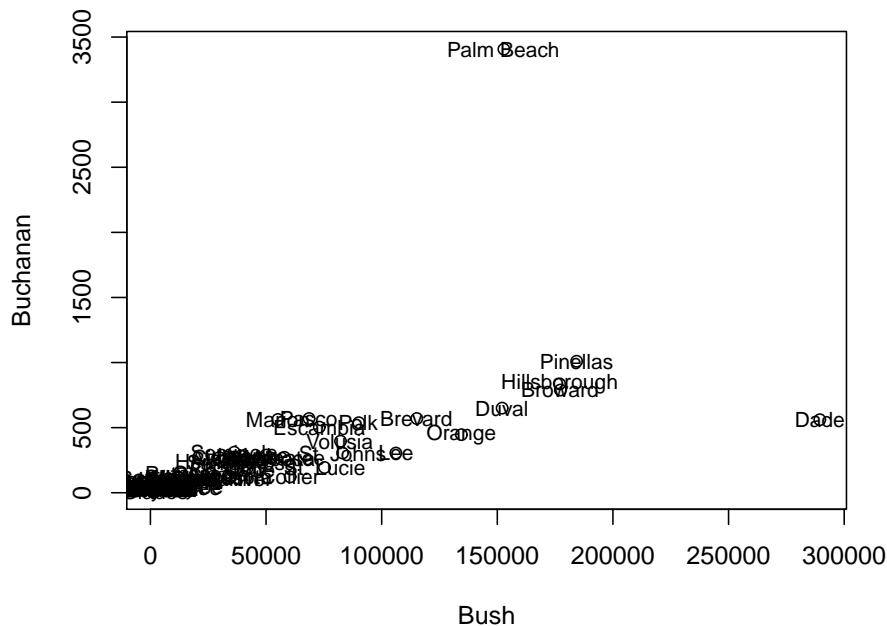


Abbildung 1: Streudiagramm zur US-Präsidentenwahl in Florida 2000.

Zur Schätzung der Buchanan-Stimmen aus den Bush-Stimmen erzeugen wir zunächst einen Datensatz ohne das County Palm Beach.

R

```
> fl_ohne_PalmBeach <- subset(fl2000, County != "Palm Beach")
> detach(fl2000)
> attach(fl_ohne_PalmBeach)
```

Mit dem reduzierten Datensatz berechnen wir über eine Einfachregression die erwarteten Stimmen für Buchanan aus den Stimmen für Bush.

R

```
> linmodfl_opb <- lm(Buchanan ~ Bush)
> coefficients(summary(linmodfl_opb))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	65.573496362	1.733043e+01	3.783721	3.427131e-04
Bush	0.003481898	2.500903e-04	13.922562	4.916245e-21

Die Beziehung zwischen den beiden Variablen ist hoch signifikant. Zur Berechnung der erwarteten Buchanan-Stimmen im County Palm Beach aus den Stimmen für Bush muss ein Dataframe mit Variablen, wie sie auch im linearen Modell vorkamen (also mit `Bush` und `Buchanan`), generiert werden. In die Variable `Bush` schreiben wir den Wert der erklärenden Variablen für die Prognose, also die Anzahl Stimmen für Bush in Palm Beach. Die Variable `Buchanan` muss im Dataframe vorkommen; ihr Wert ist egal, er wird ja erst durch die Prognose bestimmt (wir setzen sie einfach auf 0).

R

```
> predict(linmodfl_opb, newdata = data.frame(Bush = BushPB, Buchanan = 0))
```

1
597.7677

Die aus den Bush-Stimmen erwartete Anzahl an Stimmen beträgt also nicht ganz 600, erhalten hat Buchanan in Palm Beach 3407. Dieser große Unterschied ist nicht auf spezielle Beziehungen von Buchanan zu diesem County zurückzuführen (Geburtsort oder Wohnsitz in diesem County, ..), sondern auf den nur in diesem County eingesetzten Abstimmungsmodus (Wahlzettel und Stanzmaschine).

Zum Schluss lösen wir die Verbindung zum Datensatz wieder auf:

R

```
> detach(fl_ohne_PalmBeach)
```