Applied Statistical Regression
AS 2013 – Week 13

Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, December 9, 2013

Practical Example

With this example taken from the lecturer's research, we illustrate the pro's and con's of working with logistic vs. binomial regression, i.e. grouped vs. non-grouped data

CHURN	REGION	GENDER	AGE	TENURE	PRODUCT
1	D-CH	male	65	84	PH + INET + TV
1	F-CH	female	45	34	INET + TV
1	F-CH	female	68	52	INET + TV
1	D-CH	female		102	INET
1	D-CH	male	45	21	TV
1	D-CH	male	43	63	PH + INET + TV
1	I-CH	male	28	47	TV

Applied Statistical Regression
AS 2013 – Week 13

Practical Example

Goal: understanding *churn*, i.e. end of contract

Model: *churn* ~ *region* + *gender* + *age* + *tenure* + *product*

The data per se are non-grouped, with millions of observations. But in this problem, it **pays off to work with grouped data**. The main advantages when doing so are:

- Dealing with missing values in *age* and *tenure*: we do not lose any observations when factorizing these two variables.
- Instead of millions of rows, the design matrix is reduced to just 885 rows. This speeds up the computing tremendously.
- Much better inference and residual analysis is possible!

Aggregating the Data in R

```
## Aggregating the data
> gdat <- aggregate(dat$churn,by=list(dat$region, dat$sex,
                    dat$age.group, dat$dauer.group,
                    dat$produkt),table)

## Excerpt of the data
> gdat[c(34, 92, 122, 588),]
    region    sex     age     dauer   produkt churn.no churn.yes
34     F-CH   male Missing    [0,24]      PHON       53         8
92     F-CH   male (45,60]  (72,180]      PHON       50         6
122    F-CH female (30,45]    [0,24]        TV      826       194
588    F-CH female (45,60]  (72,180]   INET+TV      103        14
```

→ Now, there are $3 \cdot 3 \cdot 6 \cdot 3 \cdot 7 = 1134$ groups, of which only 885 are populated. We will now fit a binomial regression model using only the main effects (i.e. without any interaction terms).

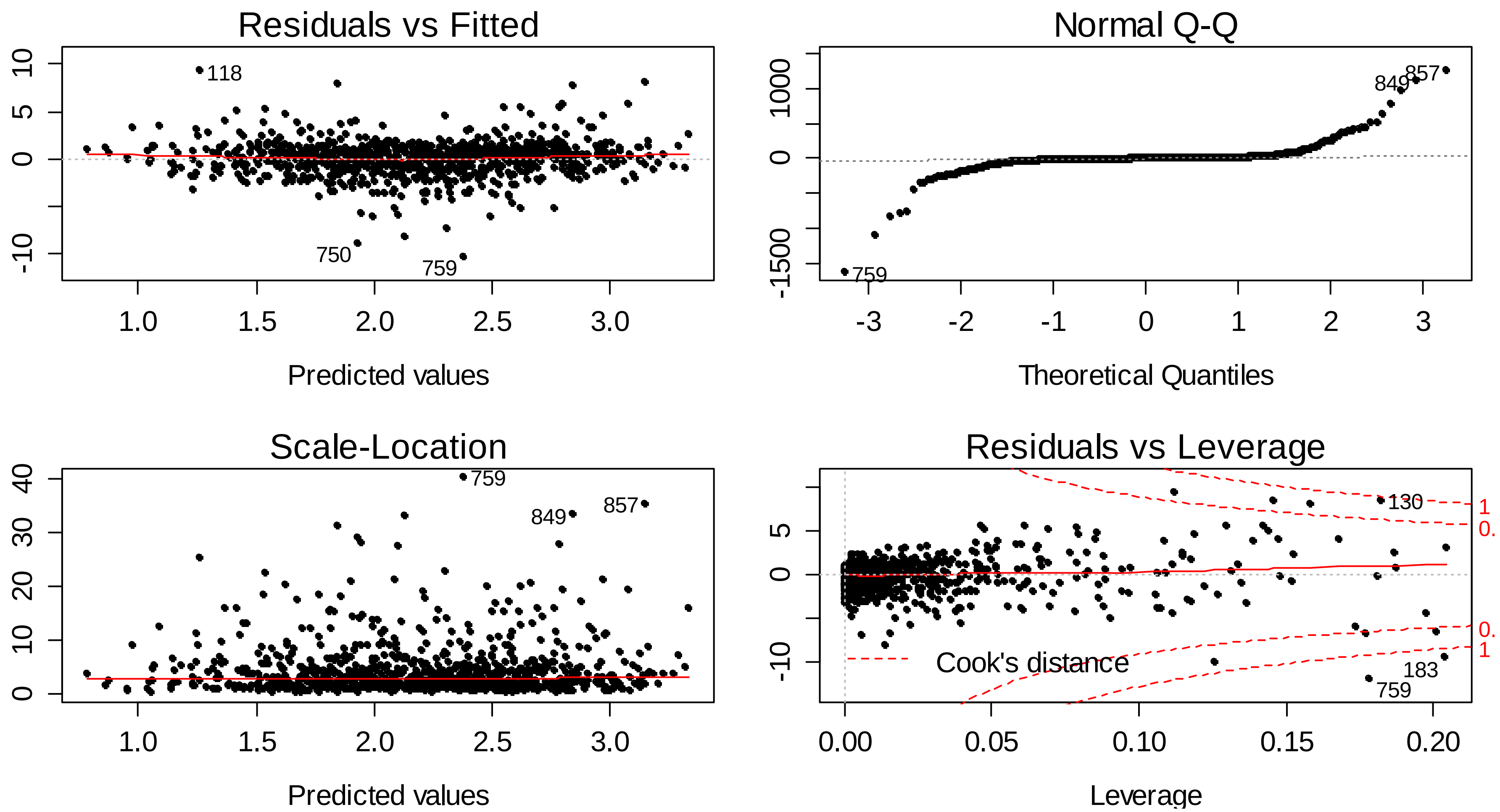
Summary Output

```
> drop1(fit, test="Chisq")

Model: churn ~ region + sex + age + dauer + produkt

        Df Deviance     AIC    LRT  Pr(>Chi)
<none>       2866.6  6254.7
region   2   3212.0  6596.1  345.4 < 2.2e-16 ***
sex      2   3344.4  6728.5  477.8 < 2.2e-16 ***
age      5   6745.2 10123.3 3878.6 < 2.2e-16 ***
dauer    2   4172.9  7557.0 1306.3 < 2.2e-16 ***
produkt  6  10718.3 14094.4 7851.7 < 2.2e-16 ***
---
Null deviance: 19369.7  on 884  degrees of freedom
Residual deviance:  2866.6  on 867  degrees of freedom
```

→ Very strong overdispersion, the model does not fit well!

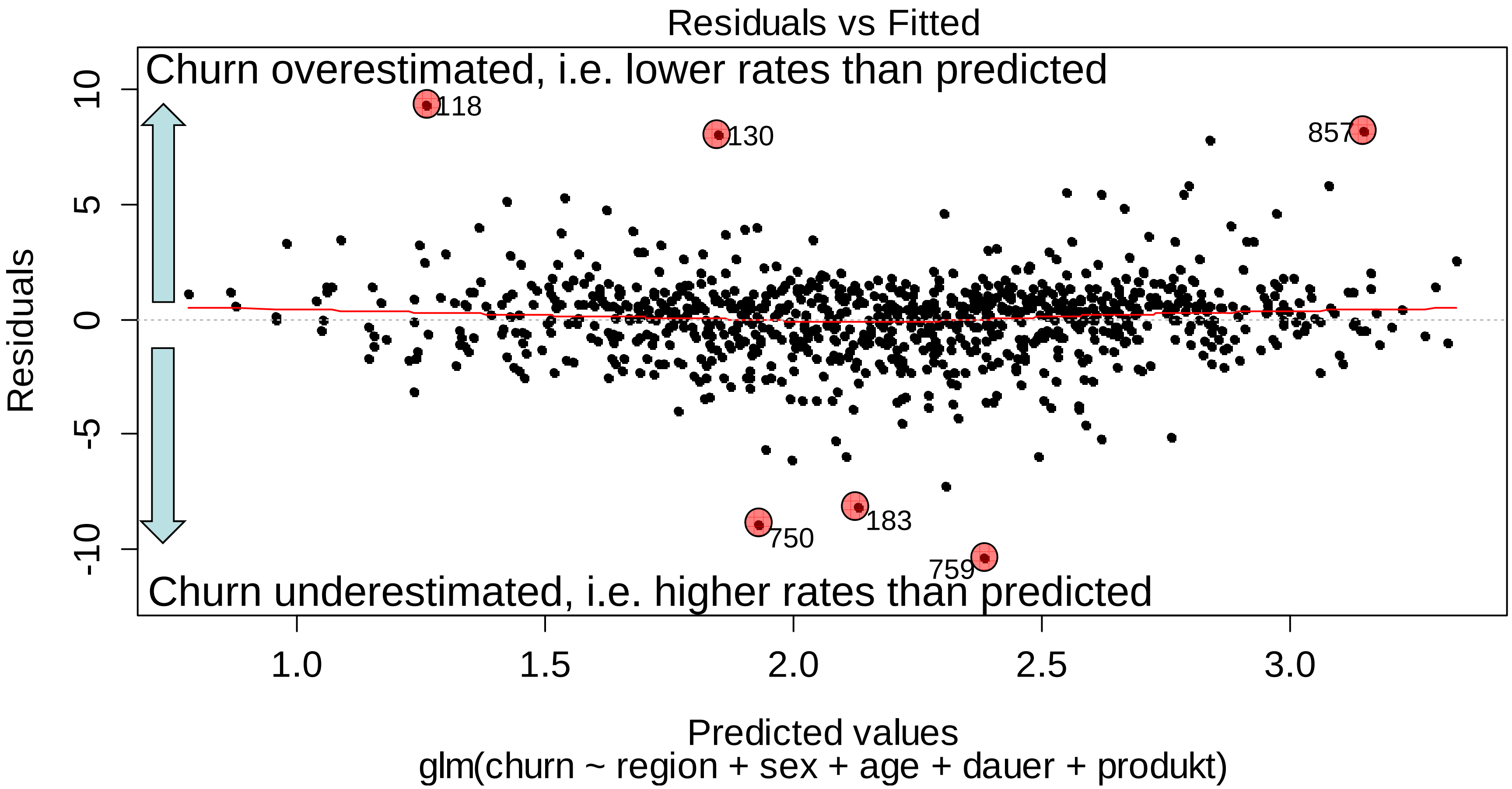
Applied Statistical Regression
AS 2013 – Week 13

Model Diagnostics



Applied Statistical Regression
AS 2013 – Week 13

Detail: Residuals vs. Predicted



Discussion of the Practical Example

The analysis of grouped data shows that we have a very incomplete understanding of the churn mechanics. There are groups for which the churn probability is very strongly over- or underestimated. All-in-all, the goodness-of-fit test for our binomial model is rejected.

What to do?

- Use more and/or better predictors for *churn*.
- If not available, try to work with interaction terms.
- Using a dispersion parameter doesn't help for prediction!
- Models can/should also be evaluated using cross validation.