Applied Statistical Regression
HS 2010 – Week 09

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, November 22, 2010

Extending the Linear Model

What is the problem?

→ **Want to model a binary response, or a proportion!**

- Variance will not be equal
- Values beyond 0/1, or beyond [0,1] will result

We need some additional techniques which can deal with these types of situations.

Depending on how the response variable is, there are several different approaches.

Logistic Regression

Example:

In human medicine, we are often interested in the question for how much „dose“ of a medication we have an effect, i.e. a reduction in pain or symptoms.

Data:

Patients, where each obtains some „dose“ and either has a reduction (1), or not (0).

There may be some further predictors such as age, sex, …

Simple Statistical Model

- A statistical model for this situation takes into account that for a given “dose”, we will only have an effect on some of the subjects, but not on all of them.

- We are thus trying to model the relation between the binary response and a number of predictors.

The ***simplest approach*** is:

$$P(Y_i = 1) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta x_{ip}$$

→ this will lead to probabilities beyond the interval of [0,1].

A Better Model

- We obtain a better model if we transform the response variable to a scale that ranges from minus to plus infinity.

- Usual choice is the so-called logit transformation:

$$p \mapsto \ln(p/(1-p))$$

We obtain the logistic regression model:

$$\log\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right)=\beta_0+\beta_1 x_{i1}+\ldots+\beta x_{ip}$$

→ all fitted values are within [0,1].

Poisson Regression

What are predictors for the locations of starfish?

→ analyze the number of starfish at several locations, for which we also have some covariates such as water temperature, ...

→ the response variable is a count. The simplest model for this is a Poisson distribution.

We assume that the parameter λ_i at location i depends in a linearly on the covariates:

$$Y_i \sim Pois(\lambda_i)\text{, where } \lambda_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

Log-Linear Models

Question:

Prediction of a nominal response variable

Example:

Which party does a person favor, depending on covariates such as education, age, sex, region, …

→ such data can be summarized with contingency tables

→ and they can be modeled using log-linear models

Applied Statistical Regression
HS 2010 – Week 09

Generalized Linear Models

What is it?

- General framework for regression type modeling
- Many different response types are allowed
- Notion: the expected value of the response has a monotone relation to a linear combination of the predictors.

$$E[Y_i] = g(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})$$

- Some further requirements on variance and density of Y

→ **may seem complicated, but is very powerful!**

Applied Statistical Regression
HS 2010 – Week 09

Binary Logistic Regression

What is it?

- Response $Y_i \in \{0,1\}$

What do we need to take care of?

- Formulation of the model
- Estimation
- Inference
- Model diagnostics
- Model choice

Example

Premature Birth, by Hubbard (1986)

$Y_i \in \{0,1\}$ survival (1) /death (0) after premature birth.

Predictors:

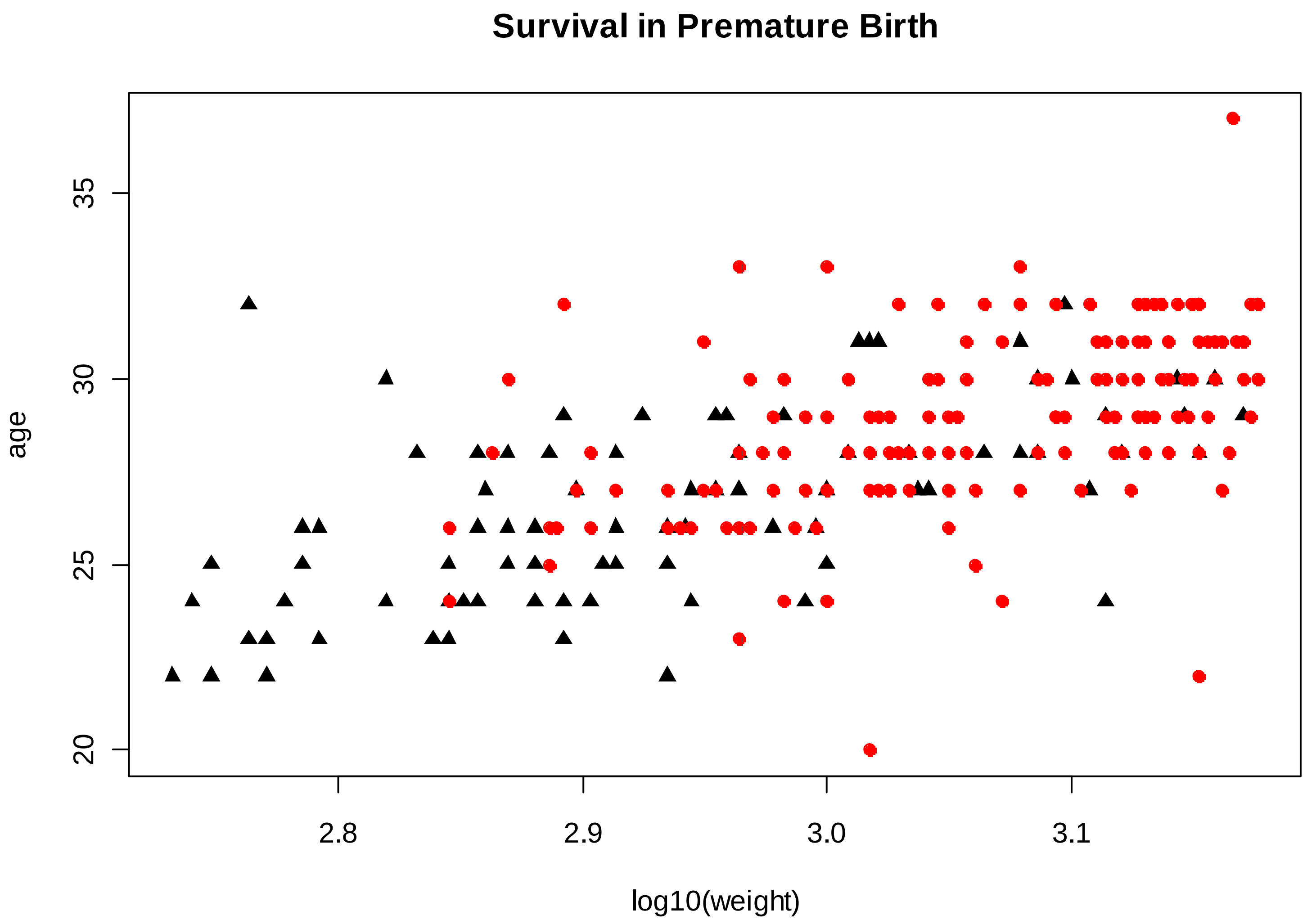
- weight (in grams) at birth
- age at birth (in weeks of pregnancy)
- apgar scores (vital function after 1 and 5 min)
- pH-value of the blood (breathing)

Observations:

- there are 247 instances

Applied Statistical Regression
HS 2010 – Week 09

Example



Applied Statistical Regression
HS 2010 – Week 09

Logistic Regression Model

- $Y_i \in \{0,1\}$ has a Bernoulli distribution.
- The parameter of this distribution is π_i, the success rate

Now please note that:

$$\pi_i = P(Y_i = 1) = E[Y_i]$$

→ the most powerful notion of the logistic regression model is to see it as a model where we try to find a relation between the expected value of Y_i and the predictors!

Important: $P(Y_i = 1) = \beta_0 + \beta_1 x_{i1} + ... + \beta x_{ip}$ is no good here!

Logit Transformation

Goal: mapping from $[0,1] \mapsto (-\infty,+\infty)$

Logit transformation: $g(\pi) = \log\left(\dfrac{\pi}{1-\pi}\right)$

Interpretation: Probabilities are mapped to log-odds ratios which can then be modeled using a linear relation.

$$\log\left(\frac{P(Y_i = 1)}{1-P(Y_i = 1)}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta x_{ip}$$

→ where is the error term?

Some Remarks

- For estimating the regression coefficients, the observations need to be independent

- There is no restriction for the predictors. They can be continuous, categorical, transformed, interactions, …

- $\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ is called the linear predictor

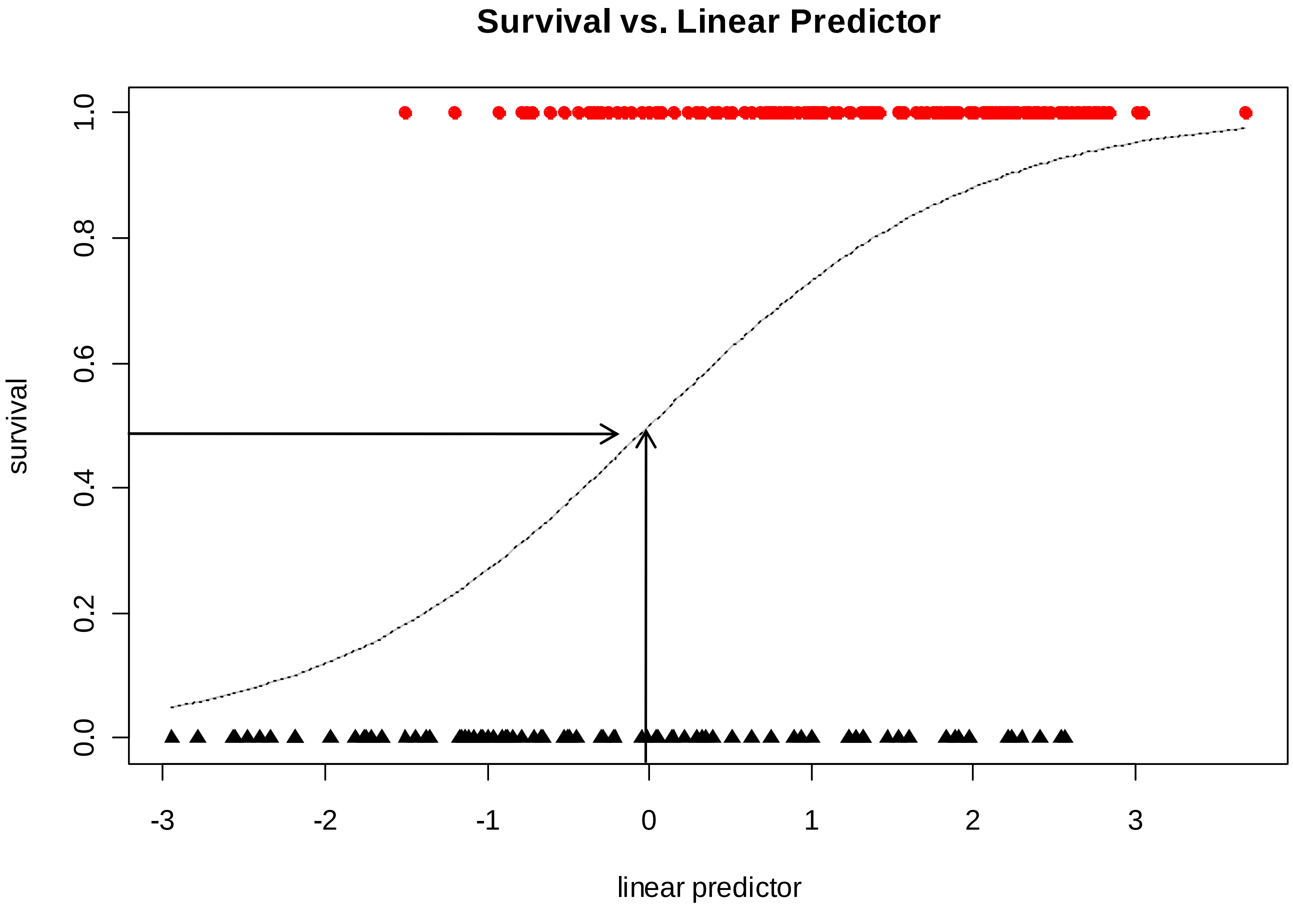
- $g(\cdot)$ is the link function, mapping between $E[Y_i]$ and η_i

- **There are other (less important) link functions:**
  - probit link
  - c-log-log link

Applied Statistical Regression
HS 2010 – Week 09

Survival vs. Linear Predictor

- $g\left(P(Y=1 \mid \log_{10}(weight), age)\right) = -33.97 + 10.17 \cdot \log_{10}(weight) + 0.14 \cdot age$



Estimation

Multiple linear regression:

→ minimize sum of squared residuals!

can be solved in closed form

Logistic regression:

→ maximum likelihood approach!

leads to a non-linear equation system that needs to be solved with an iterative approach by weighted multiple linear regressions.

Important:

→ seems like a very different paradigm, but is it?

Interpretation of the Coefficients

→ see blackboard…

Inference

```
> summary(glm(survival ~ I(log10(weight)) + age,
              family  = "binomial", data = baby)

Deviance Residuals: ...

Coefficients:     Estimate Std. Error z value Pr(>|z|)
(Intercept)      -33.97108    4.98983  -6.808 9.89e-12 ***
I(log10(weight))  10.16846    1.88160   5.404 6.51e-08 ***
age                0.14742    0.07427   1.985   0.0472 *
---
    Null deviance: 319.28  on 246  degrees of freedom
Residual deviance: 235.94  on 244  degrees of freedom
AIC: 241.94
```

Individual Parameter Tests

Multiple Linear Regression:

Gaussian errors → $\hat{\beta}_j$ are normally distributed

Logistic Regression:

There are no errors, variability arises from Bernoulli distribution

The regression coefficients $\hat{\beta}_j$ are only approximately normally distributed with a covariance matrix V that can be derived from the coefficients.

Hence: $$Z = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}_{jj}}} \sim N(0,1)$$

Goodness-of-fit

Multiple Linear Regression:

Sum of Squared Residuals

Logistic Regression:

Residual Deviance

$$D(y, \hat{\pi}) = -2\sum_i (y_i \log(\hat{\pi}_i) + (1 - y_i)\log(1 - \hat{\pi}_i))$$

- based on the log-likelihood
- in principle: comparison against fully saturated model

Applied Statistical Regression
HS 2010 – Week 09

Comparing Nested Models

Model 1: small model S, with q parameters

Model 2: big model B, with p parameters

Null hypothesis and test statistic:

$$H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$$

$$2\left(ll^{(B)} - ll^{(S)}\right) = D\left(y, \hat{\pi}^{(S)}\right) - D\left(y, \hat{\pi}^{(B)}\right)$$

Distribution of the test statistic:

$$D^{(S)} - D^{(B)} \sim \chi^2_{p-q}$$

Example with drop1()

```
> drop1(fit, test="Chisq")
Single term deletions
Model: survival ~ I(log10(weight)) + age
                 Df Deviance    AIC    LRT   Pr(Chi)
<none>                235.94 241.94
I(log10(weight))  1   270.19 274.19 34.247 4.855e-09 ***
age               1   239.89 243.89  3.948   0.04694 *
```

Question:

- where is the difference to the summary output?
- it exists, though it‘s not obvious and asymptotically vanishes

AIC and Variable Selection

General remark:

All comparison between models of different size can also be done using the AIC criterion. Not only in logistic regression, but also here.

The criterion:

$$AIC = D(y_i, \hat{\pi}) + 2p$$

Variable selection:

- stepwise approaches as with multiple linear regression
- factor variables need to be treated the right way!

Null Deviance

Smallest model:

- The smallest model is without predictors, only with intercept
- Fitted values will all be equal to $\hat{\pi}_0$
- Our best fit (F) and the smallest model (0) are nested

A global test:

$$2\left(ll^{(F)} - ll^{(0)}\right) = D\left(y, \hat{\pi}^{(0)}\right) - D\left(y, \hat{\pi}^{(F)}\right)$$

Example:

```
Null deviance: 319.28  on 246  degrees of freedom
Residual deviance: 235.94  on 244  degrees of freedom
```

Applied Statistical Regression
HS 2010 – Week 09

Model Diagnostics

Diagnostics are:

- as important with logistic regression as they are with multiple linear regression models
- again based on differences between fitted & observed values

→ we now have to take into account that the variances are not equal for the different instances.

→ we have to come up with novel types of residuals:

Pearson and **Deviance residuals**

Applied Statistical Regression
HS 2010 – Week 09

Pearson Residuals

Take the difference between observed an fitted value and divide by an estimate of the standard deviation:

$$R_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$$

→ R_i^2 is the contribution of the *ith* observation to the Pearson statistic for model comparison.

→ It is important to note that Pearson residuals exceeding a value of two in absolute value warrant a closer look

Applied Statistical Regression
HS 2010 – Week 09

Deviance Residuals

Take the contribution of the i*th* observation to the log-likelihood, i.e. the chi-square statistic for model comparison.

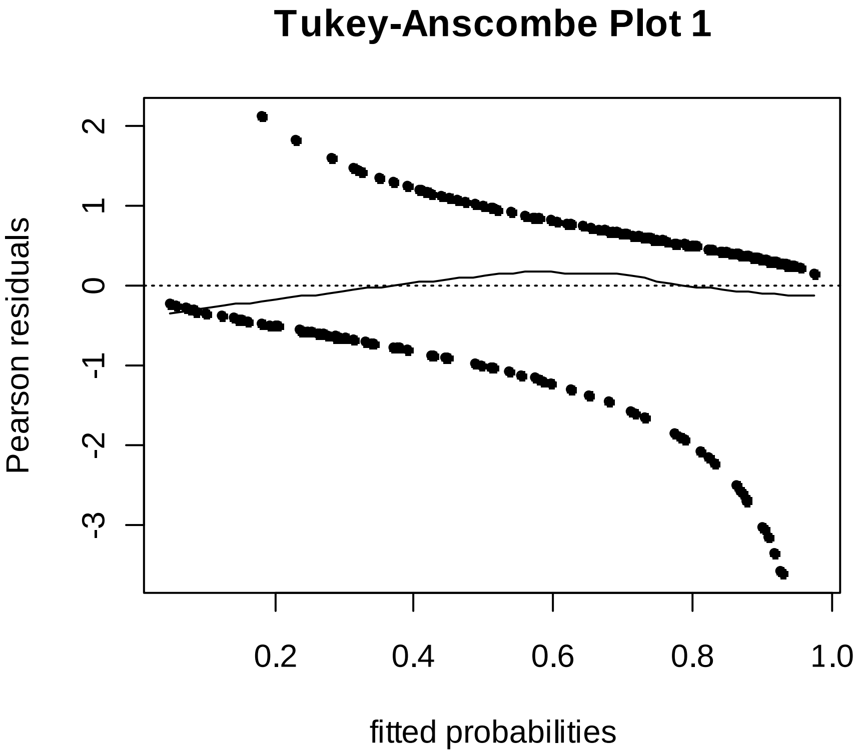
$$d_i = \left(y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)\right)$$

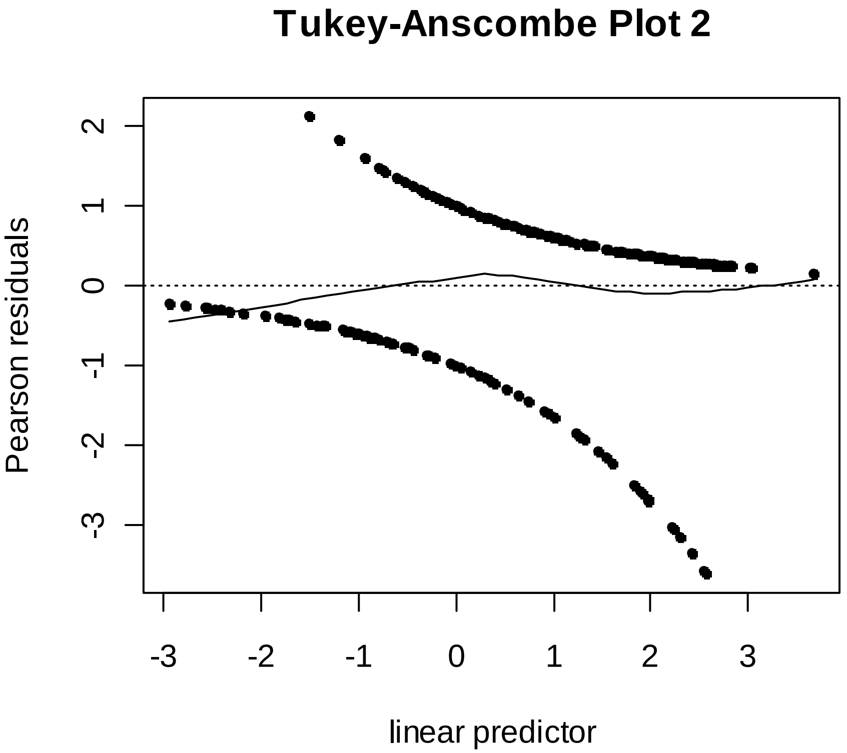
For obtaining a well interpretable residual, we take the square root and the sign of the difference between true and fitted value:

$$D_i = sign(y_i - \hat{\pi}_i) \cdot \sqrt{d_i}$$

→ It is important to note that Pearson residuals exceeding a value of two in absolute value warrant a closer look

Tukey-Anscombe Plot





Tukey-Anscombe Plot

The Tukey-Anscombe plots in R are not perfect. Better use:

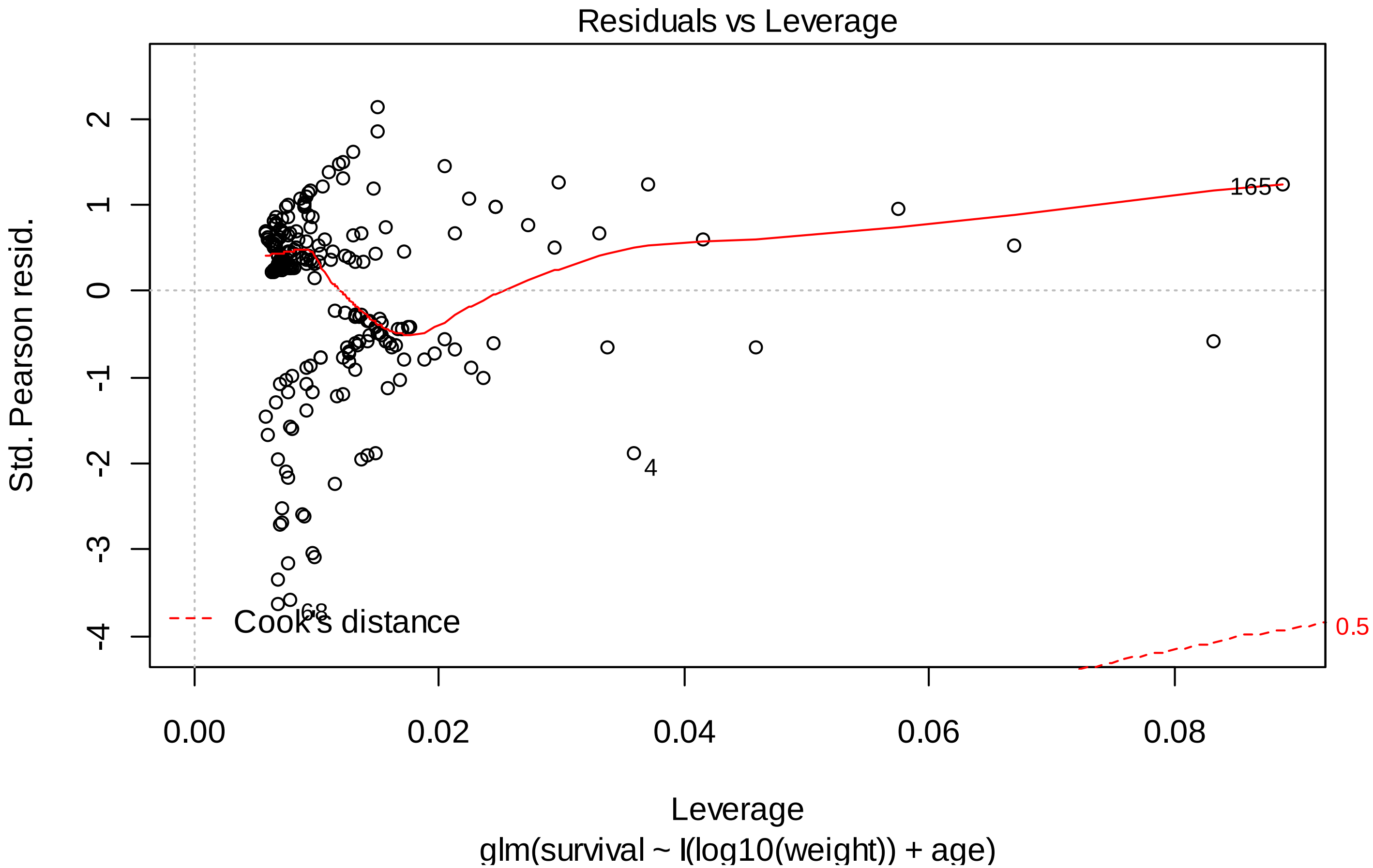
```
xx <- predict(fit, type="response")
yy <- residuals(fit, type="pearson")
scatter.smooth(xx, yy, family="gaussian", pch=20)
abline(h=0, lty=3)
```

Reasons:

- using a non-robust smoother is a must
- different types of residuals can be used
- on the x-axis: probs or linear predictor

Applied Statistical Regression
HS 2010 – Week 09

More Diagnostics


Residuals vs Leverage
Std. Pearson resid.
Cook's distance
Leverage
glm(survival ~ I(log10(weight)) + age)