Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

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 2014

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Extending the Linear Model

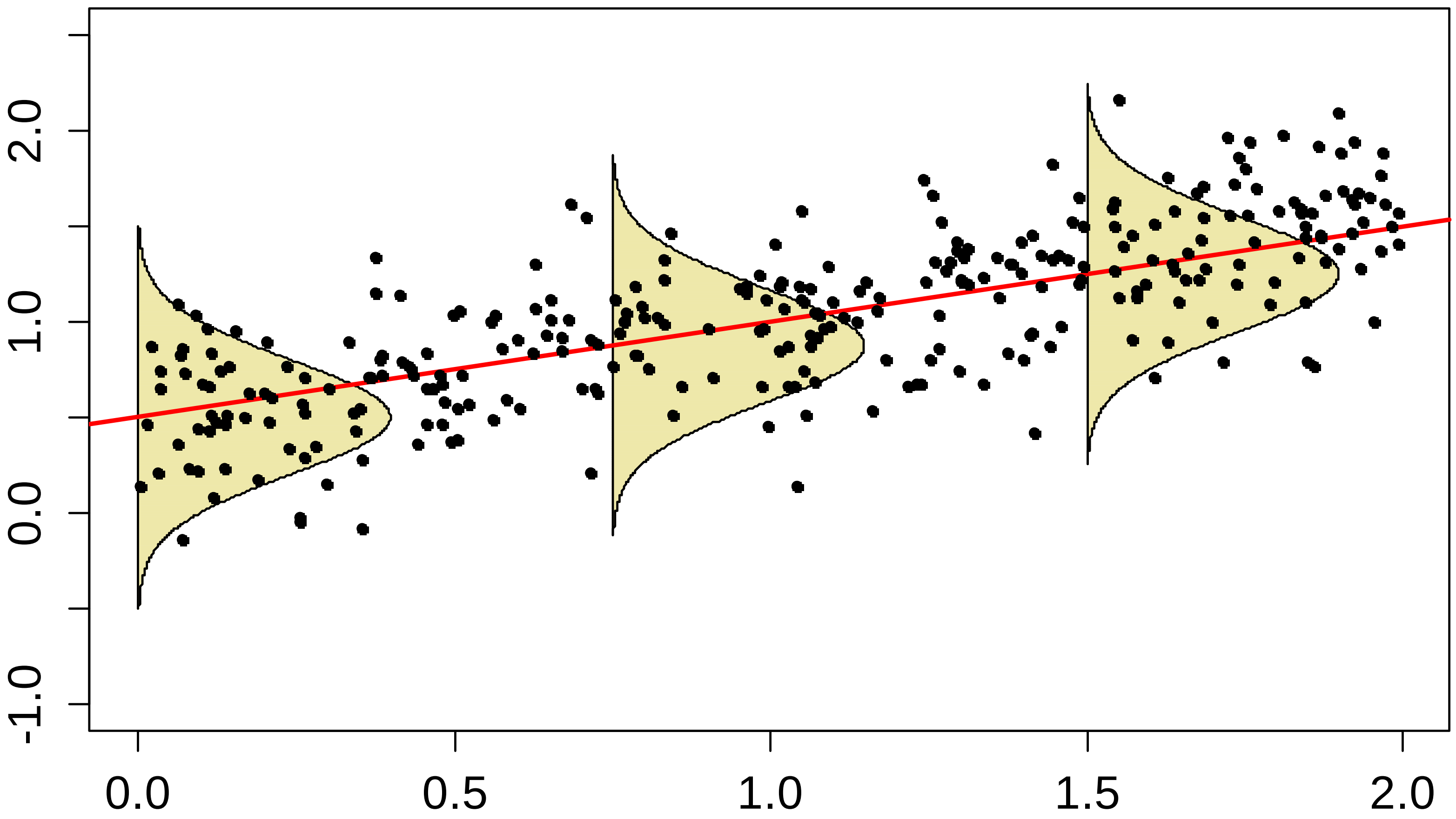
→ **So far, the response y_i was a continuous random variable with infinite range, where the conditional distribution was a Gaussian, i.e. $y_i \mid X_i \sim N(\hat{y}_i, \sigma_E^2)$, see next slide.**

→ If the task is modeling binary, binary or multinomial response (i.e. probabilities or proportions) or a count, this is not doable correctly with the methods that were discussed yet.

→ We will present some additional techniques which implement linear modeling for these different types of responses. As we will see, there is a generic framework that incorporates all of these, as well as multiple linear regression.

Conditional Gaussian Distribution

Linear Regression with Gaussian Distributions



Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Binary Response / Logistic Regression

What is the question?

In toxicological studies, one tries to infer wheter a lab mouse survives when it is given a particular dose of poisonous matter.

In human medicine, one is often interested in the question how much of a drug is required to see an effect, i.e. pain reduction.

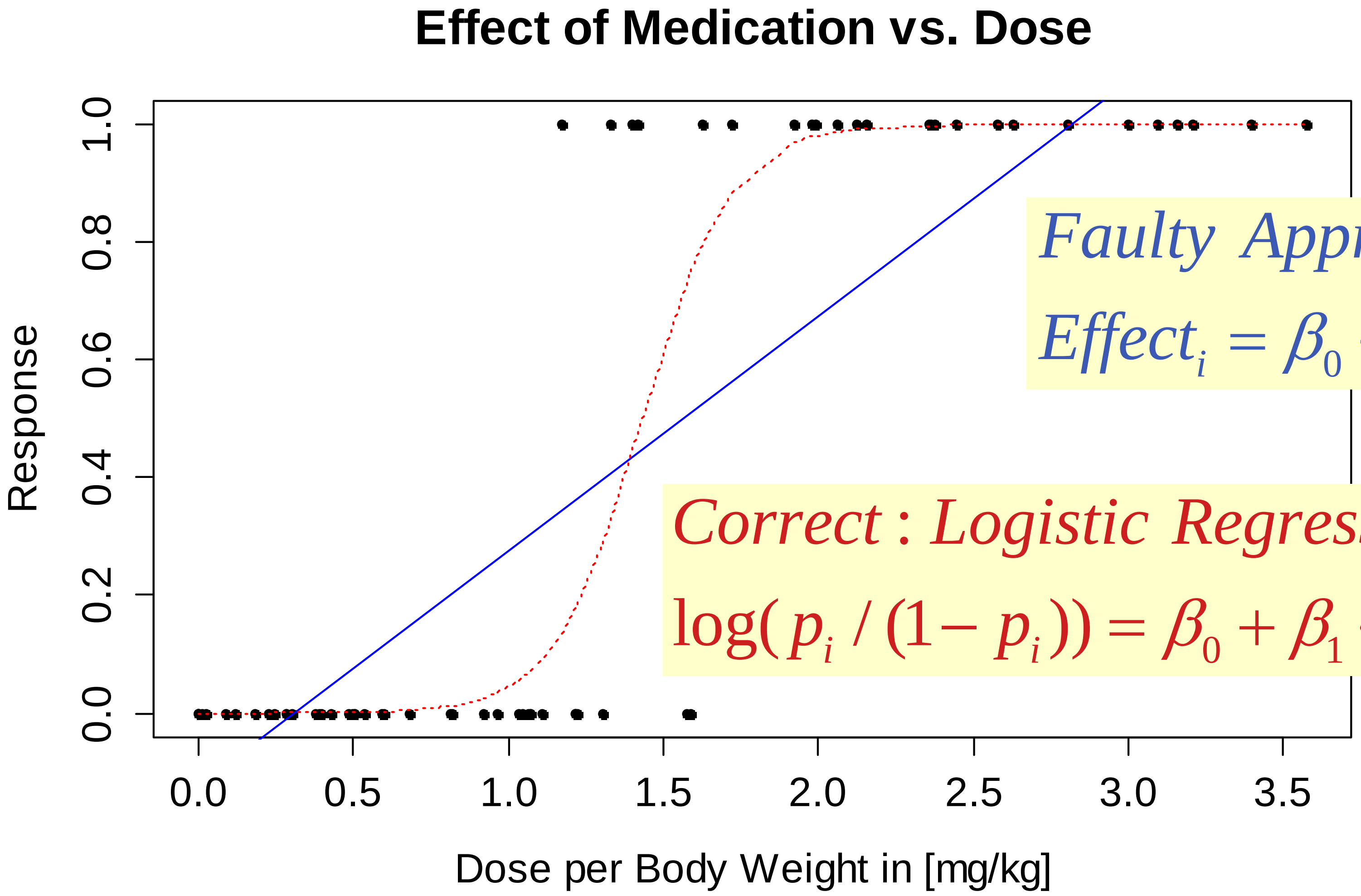
Mathematics:

→ The response variable $y_i \in \{0,1\}$ is binary

→ The conditional distribution $y_i \mid X_i \sim Bernoulli(p_i)$

→ The fitted value is the expectation of the above conditional distribution, and hence the probability of death/survival p_i.

Binary Response / Logistic Regression



Faulty Approach:

$$Effect_i = \beta_0 + \beta_1 \cdot Dose_i + E_i$$

Correct: Logistic Regression

$$\log(p_i / (1 - p_i)) = \beta_0 + \beta_1 \cdot Dose_i$$

Applied Statistical Regression
AS 2014 – Generalized Linear Modeling

Count Response / 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 assumes a Poisson as the conditional distribution.

We assume that the logged parameter λ_i at location *i* depends in a linear way on the covariates:

$$y_i \mid X_i \sim Pois(\lambda_i), \text{ where } \log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

Generalized Linear Models

What is it?

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

$$g(E[y_i \mid X_i]) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

- Some further requirements on variance and density of y

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

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Binary Response / Logistic Regression

The essence of a logistic regression model is that the response $y_i \in \{0,1\}$, the conditional distribution is $y_i \mid X_i \sim Bernoulli(p_i)$ and we model the conditional expectation $E[y_i \mid X_i] = p_i$.

What do we need to take care of?

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

Applied Statistical Regression
AS 2014 – Generalized Linear Modeling

Example

Premature Birth, by Hubbard (1986)

$y_i \in \{0,1\}$ death (0) or survival (1) 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)

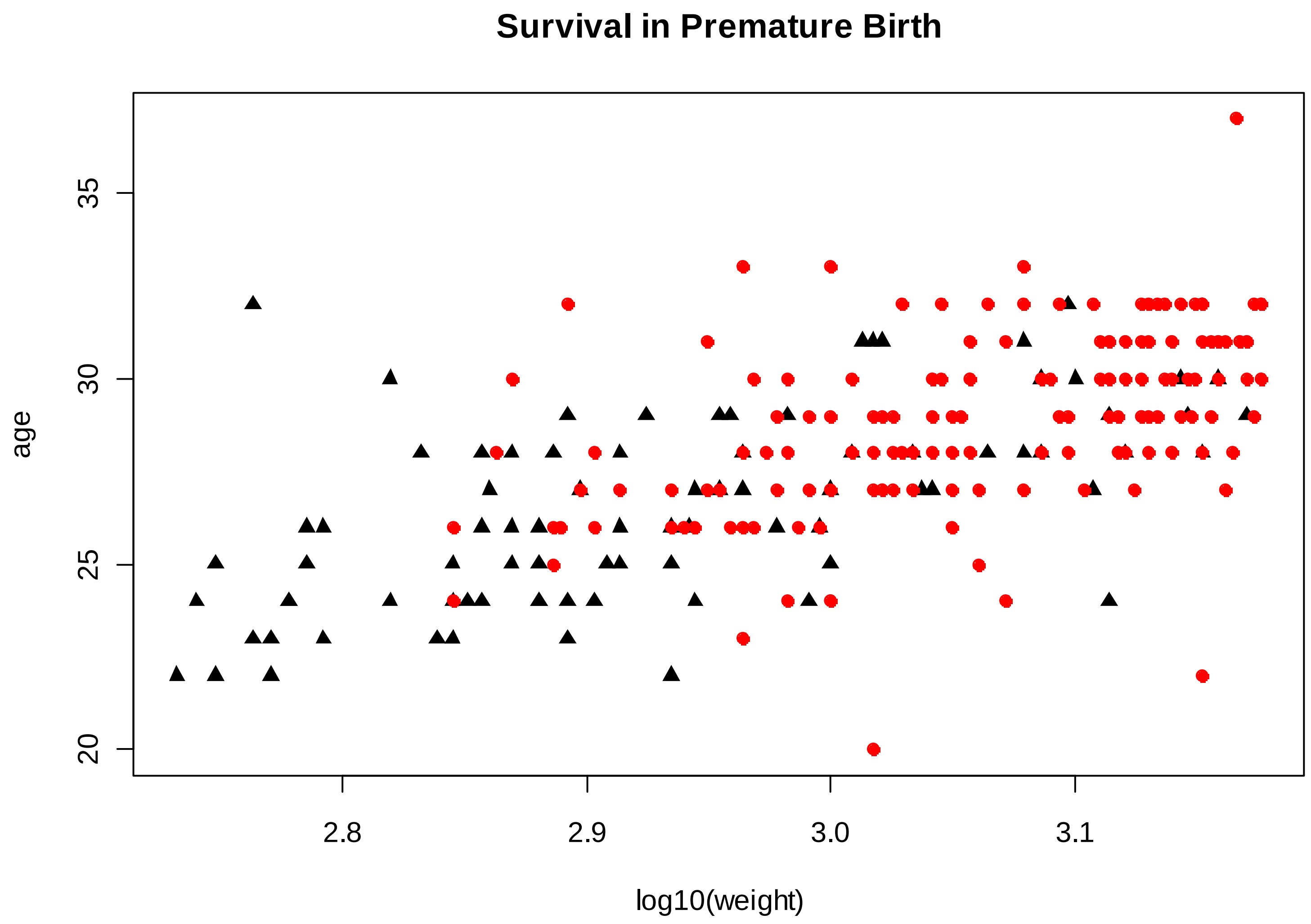
Observations:

- there are 247 instances

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Example



Logistic Regression Model

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

Now please note that:

$$p_i = P(y_i = 1 \mid X_i) = E[y_i \mid X_i]$$

→ 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 conditional expected value of y_i and the predictors!

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

Logit Transformation

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

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

Interpretation: probabilities are mapped to logged odds ("Wettverhältnisse") which can then be modeled linearly.

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta x_{ip} \Leftrightarrow p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{1+\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}$$

→ Logistic regression = describing log-odds with a linear model

→ **Can you explain why there is no error term?**

Some Remarks and Terminology

- For estimating the regression coefficients, we require the observations 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 from $E[y_i \mid X_i]$ to η_i
- **There are other (less important) link functions:**
  - *probit link*
  - *c-log-log link*

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Estimation

Simple approach: minimize $\sum_{i=1}^{n}(y_i - p_i)^2$

→ Not a good way to estimate logistic regression parameters

Maximum-Likelihood approach:

General principle: determine the regression coefficients β_j such that the likelihood of the observed data is maximized. If the cases are independent, this amounts to maximizing the log-likelihood:

$$l(\beta) = \sum_{i=1}^{n}\left(y_i \log(p_i) + (1-y_i)\log(1-p_i)\right) \text{ with } p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots)}{1+\exp(\beta_0 + \beta_1 x_{i1} + \ldots)}$$

Under mild conditions, the solution exists, but it cannot be written in closed form. Usually, the IRLS algorithm is employed.

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Estimation in R

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

Coefficients: (Intercept)  log10(weight)      age
                 -33.9711        10.1685   0.1474
```

These are the estimates for $\beta_0, \beta_1, \beta_2$. Please note that they come from a numerical optimization, thus don't ignore this lightly:

```
Warning message:
glm.fit: algorithm did not converge
```

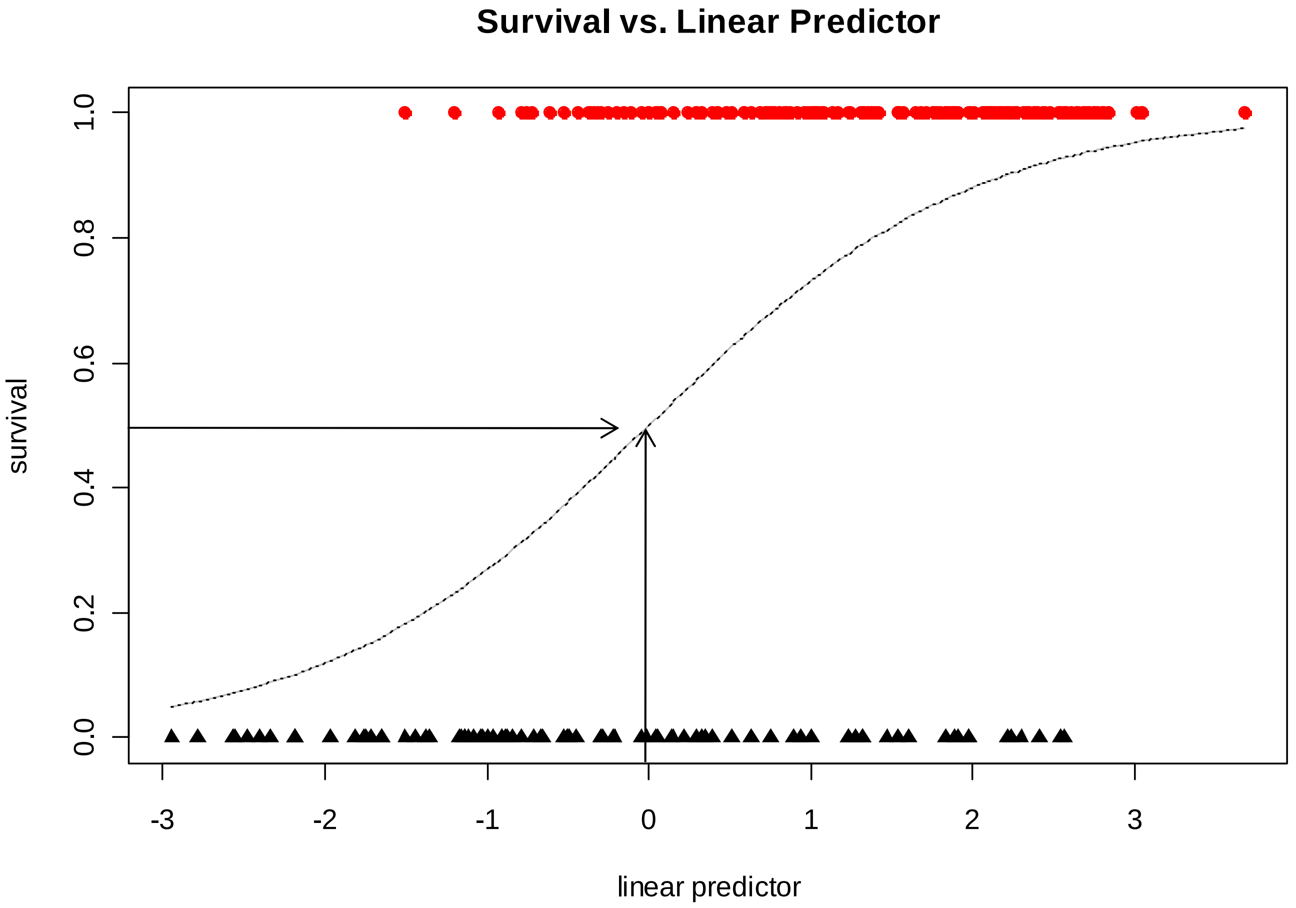
In this case, the coefficients are not trustworthy! However, this rarely happens in well posed regression problems.

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Displaying the Fit

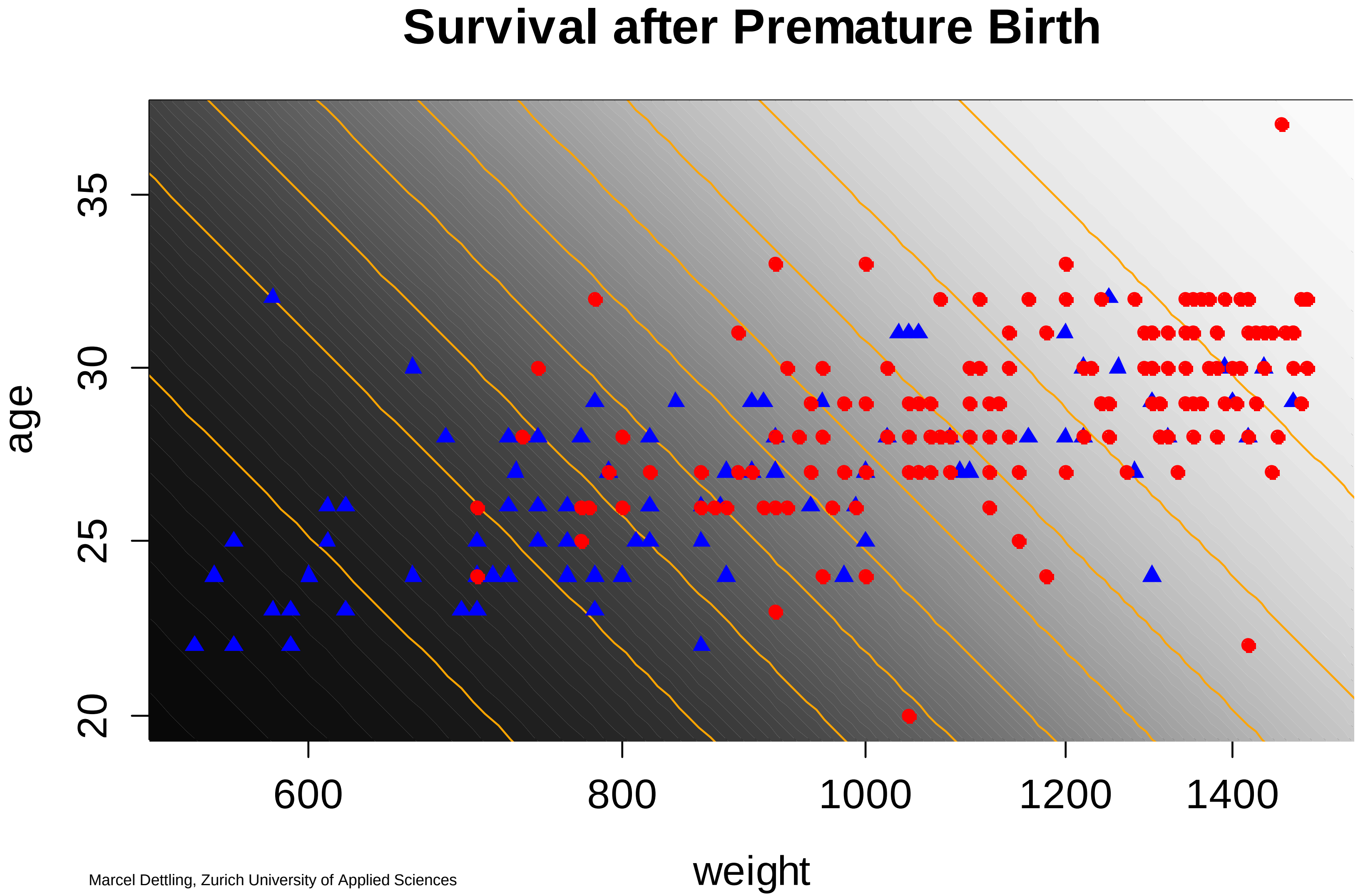
$$g\left(P(y=1 \mid \log_{10}(weight), age)\right) = -33.97 + 10.17 \cdot \log_{10}(weight) + 0.14 \cdot age$$



Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Displaying the Fit



Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Interpretation of the Coefficients

→ **see blackboard…**

Inference for Logistic Regression

While the basic concepts of inference will be familiar from multiple linear regression, various aspects will be markedly different.

Most importantly, the concept for the goodness-of-fit measure needs a second thought. We cannot work with the residuals sum of squares anymore, but employ the so-called **Residual Deviance**:

$$D(y, \hat{p}) = -2 \cdot \sum_{i=1}^{n} \left(y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i) \right)$$

Also of importance is the **Null Deviance**, which is the deviance of the simplest possible model that is built from the intercept term only. It is always lower than the Residual Deviance.

Summary Output from R

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

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 *
---
Dispersion parameter for binomial family taken to be 1
Null deviance: 319.28  on 246  degrees of freedom
Residual deviance: 235.94  on 244  degrees of freedom
Number of Fisher Scoring iterations: 4
AIC: 241.94
```

Coefficient of Determination

There is no direct analogon to the coefficient of determination in logistic regression. Some suggestions for the COD include:

Proportion of deviance explained

```
> 1-fit$dev/fit$null
[1] 0.2610193
```

A better statistic for measuring the explanatory content

$$R^2 = \frac{1-\exp((D_{res}-D_{null})/n)}{1-\exp(-D_{null}/n)} = 0.395$$

→ There are even more suggestions in the literature.

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Inference: 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

MLE-theory tells us that under mild conditions, the coeffcients $\hat{\beta}_j$ are approximately normally distributed with a covariance matrix V that can be derived from the coefficients.

Hence: If $H_0 : \beta_j = b$ then use $Z = \dfrac{\hat{\beta}_j - b}{\hat{\sigma}_{\hat{\beta}_j}} \sim N(0,1)$

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Inference: 95%-CI for the Coefficients β_j

It is straightforward to generate a hand-constructed 95%-CI:

$$\hat{\beta}_j \pm 2 \cdot se(\hat{\beta}_j) = \hat{\beta}_j \pm 2 \cdot \hat{\sigma}_{\hat{\beta}_j}$$

Rather than using the approximate 2 for the 97.5%-quantile of the Gaussian distribution, we can use the exact values. For log10(weight), we so obtain:

```
> 10.16846+qnorm(c(0.025,0.975))*1.88160
[1] 6.480592 13.856328
```

However, in R it is more convenient to use `confint()` which here uses a slighty different, more sophisticated and exact computation by interpolation of the likelihood profile traces.

Inference: Comparing Hierarchical Models

Analogon to the partial F-test in multiple linear regression

Big Model: has $(p+1)$ coefficients $\beta_0, \beta_1, ..., \beta_q, ..., \beta_p$
Small Model: has $(q+1)$ coefficients $\beta_0, \beta_1, ..., \beta_q$

Null hypothesis: $H_0 : \beta_{q+1} = \beta_{q+2} = ... = \beta_p = 0$

MLE theory suggests to use the likelihood ratio or log-likelihood difference as a test statistic. This amounts to taking the difference of the residual deviances. It asymptotically follows a Chisquare distribution with $(p-q)$ degrees of freedom:

$$2\left(ll^{Big} - ll^{Small}\right) = D\left(y, \hat{p}_{Small}\right) - D\left(y, \hat{p}_{Big}\right) \sim \chi^2_{(p-q)}$$

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Example: Global Test

Idea: Compare the actual model against the simplest possible model with only the intercept. The latter fits the overall success ratio $\hat{p}_{Null} = \sum y_i / n$ to all observations.

Since our actual model and the null model are nested, we can perform a hierarchical model comparison. In the baby survival example, there are two predictors and hence:

$$D\left(y, \hat{p}_{Null}\right) - D\left(y, \hat{p}_{Big}\right) \sim \chi_2^2$$

The two deviances are reported in the summary output.

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

Example: Global Test

A quick and simple check for rejection of $H_0 : \beta_1 = \beta_2 = 0$ is to compare the difference in deviance vs. the difference in degrees of freedom.

If $D(y, \hat{p}_{Null}) - D(y, \hat{p}_{Big}) \gg (p - q)$ then reject H_0

The exact p-valued can be computed in R by:

```
> 1-pchisq(fit$null-fit$dev, df=(fit$df.null-fit$df.res))
[1] 0
```

The p-value is (numerically) zero, hence the null hypothesis is very clearly rejected. Conjecture: there is a strongly significant contribution of `log10(weight)` and age to the odds for survival.

Using `drop1()` *for Testing*

The R function `drop1()` performs hierarchical model comparison for exclusion of one model term at a time.

```
> drop1(fit, test="Chisq")
Single term deletions
Model: survival ~ I(log10(weight)) + age
                 Df Deviance    AIC    LRT   Pr(Chi)
<none>               235.94 241.94
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

Model Diagnostics

Diagnostics are:

- in principle as important with logistic regression as they are with multiple linear regression models, but more difficult.

- again based on differences between fitted & observed values

→ we have to take into account that the variance of the response residuals $r_i = y_i - \hat{p}_i$ is non-constant.

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

Pearson and ***Deviance residuals***

Pearson Residuals

Take response residual (difference between observed and fitted value) and divide by an estimate of its standard deviation:

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

→ R_i^2 is the contribution of the *ith* observation to the Pearson statistic for model comparison (that we did not discuss).

→ It is important to note that Pearson residuals exceeding a value of two in absolute value warrant a closer look. They appear if an observation with $\hat{p}_i = 0.8$ resp. $\hat{p}_i = 0.2$ in reality has class label 0 resp. 1.

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Deviance Residuals

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

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

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

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

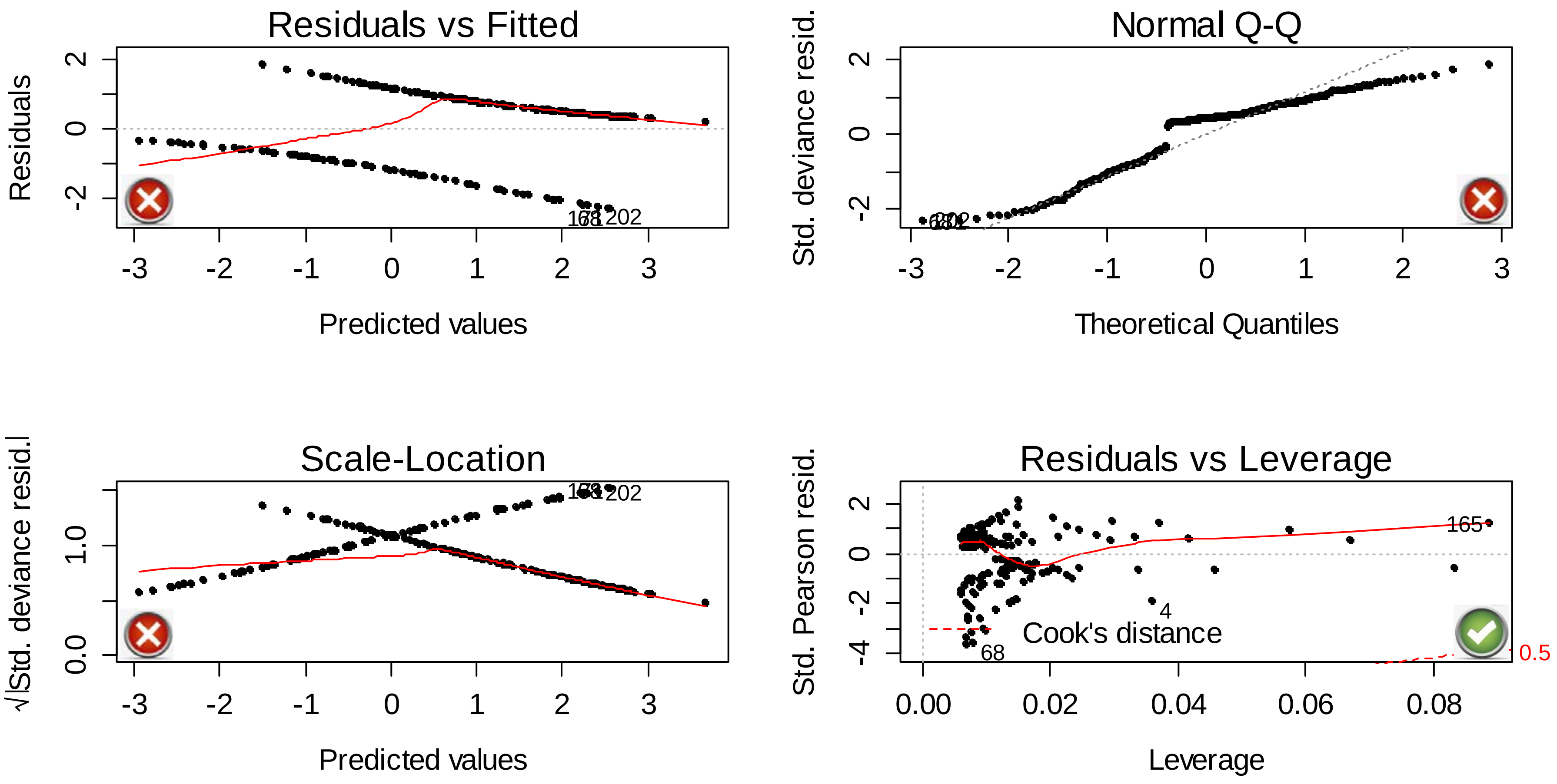
→ - *deviance residuals > 2 warrant a closer look*.

- *the distribution of the deviance residuals is not known*.

Applied Statistical Regression
AS 2014 – Generalized Linear Modeling

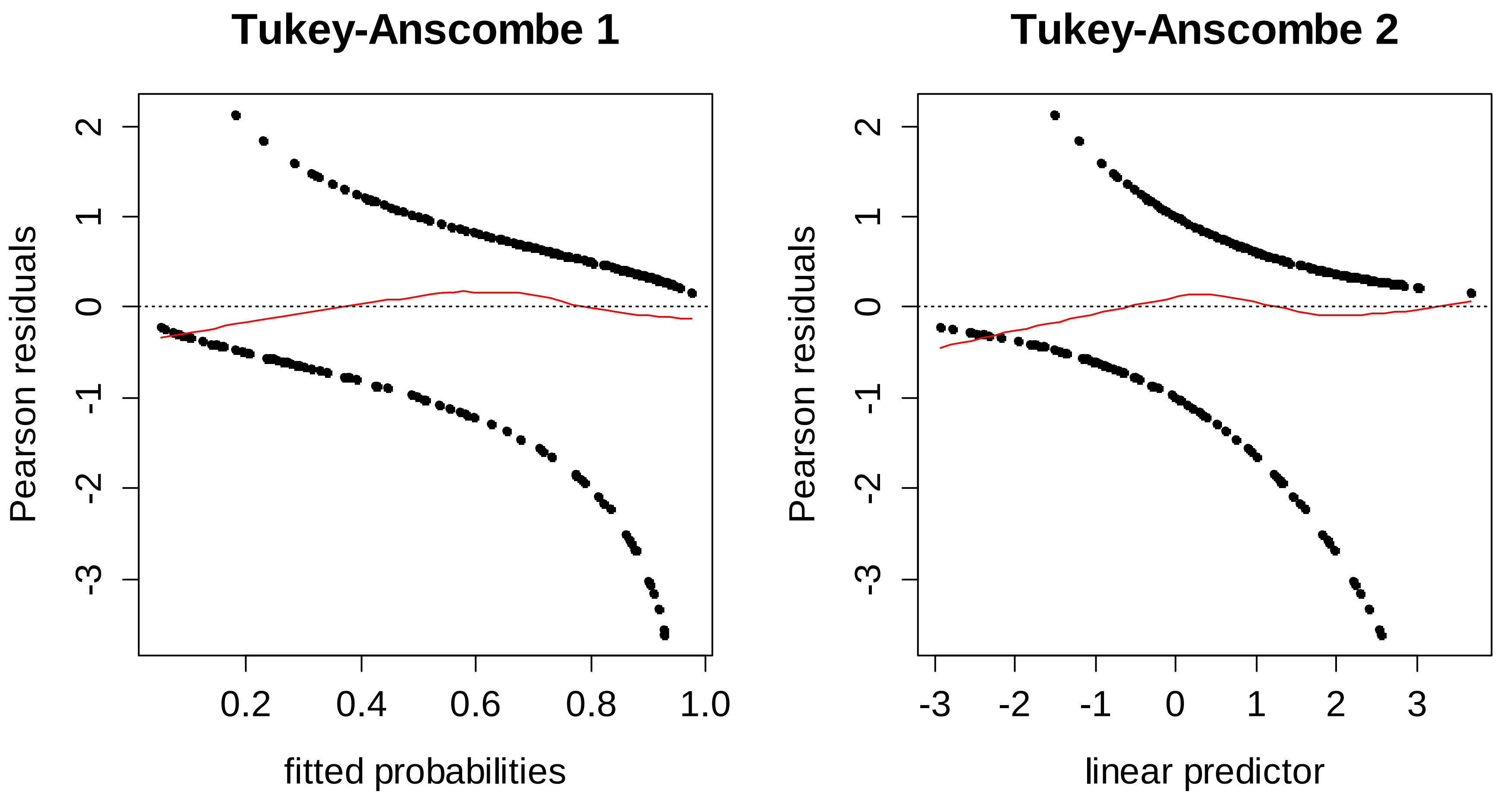
Model Diagnostics in R

The 4 standard plots are not well suited for logistic regression!!!



Applied Statistical Regression
AS 2014 – Generalized Linear Modeling

Improved Tukey-Anscombe Plots



Improved Tukey-Anscombe Plot

It is mandatory to use a non-robust smoother in the TA plot!

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

Remarks:

- On the y-axis, use *Pearson* or *Deviance* residuals
- On the x-axis, use the *linear predictor* or *probabilities*
- One can, but does not have to use studentized residuals
- The LogReg residuals do not follow a Gaussian distribution
- The LogReg residuals always lie on two curves
- Residual analysis is easier with grouped data!

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:

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

Variable selection:

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

Binomial Regression Models

Concentration in log of mg/l	Number of insects n_i	Number of killed insects y_i
0.96	50	6
1.33	48	16
1.63	46	24
2.04	49	42
2.32	50	44

→ For the number of killed insects, we have $y_i \sim Bin(n_i, p_i)$

→ We are mainly interested in the proportion of insects surviving

→ These are grouped data for which we do binomial regression. We could run a logistic regression with 243 observations instead, but the grouped data approach is more powerful!

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Model and Estimation

The goal is to find a relation:

$$p_i = P(y_i = 1 \mid X_i) \sim \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

We will again use the logit link function such that $\eta_i = g(p_i)$

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

Here, p_i is the expected value $E[y_i / n_i]$, and thus, also this model here fits within the GLM framework. The log-likelihood is:

$$l(\beta) = \sum_{i=1}^{k} \left[\log\begin{pmatrix} n_i \\ y_i \end{pmatrix} + n_i y_i \log(p_i) + n_i(1-y_i)\log(1-p_i) \right]$$

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Fitting with R

We need to generate a two-column matrix where the first contains the “successes” and the second contains the “failures”

```
> killsurv
     killed surviv
[1,]      6     44
[2,]     16     32
[3,]     24     22
[4,]     42      7
[5,]     44      6
```

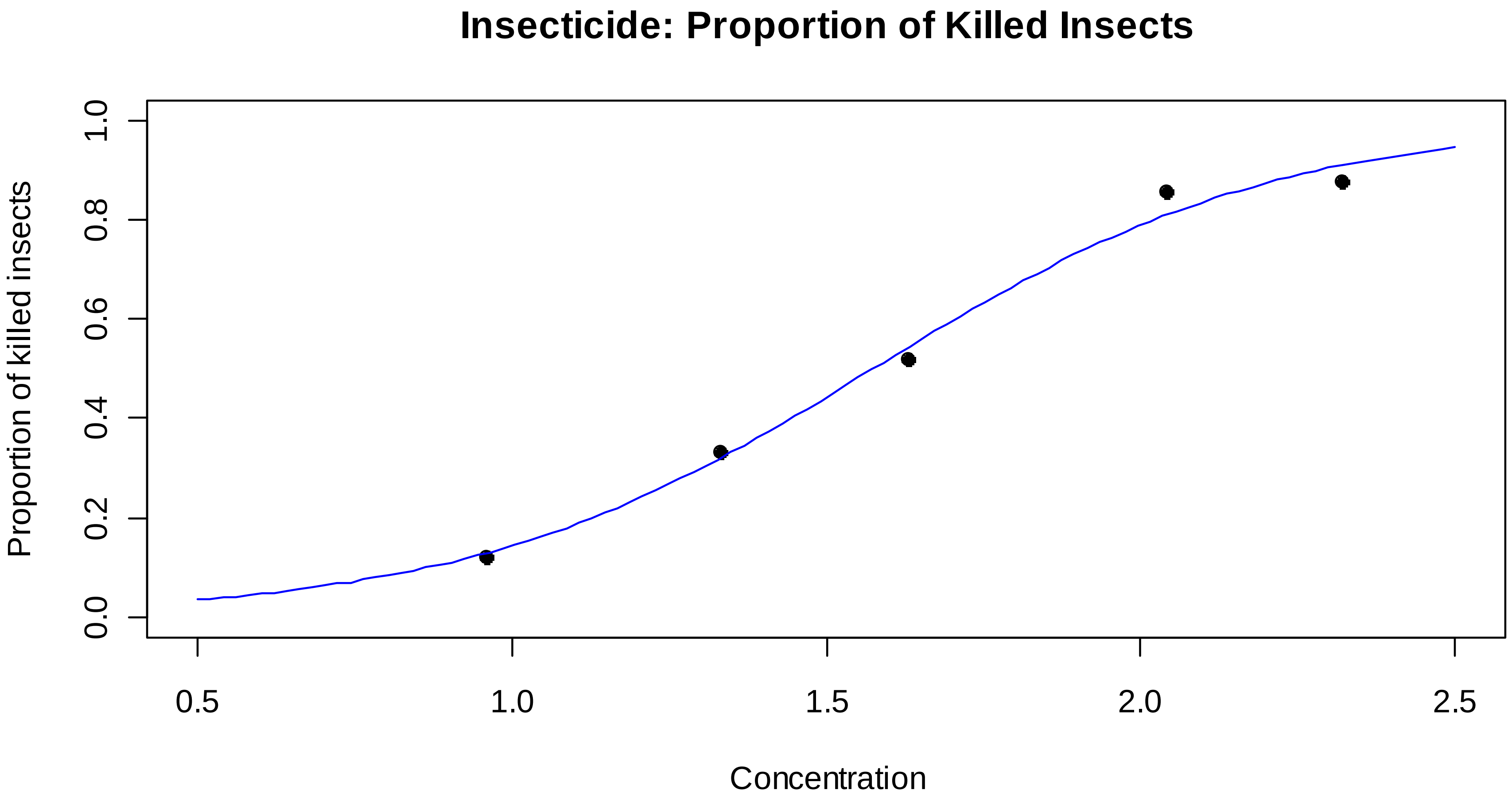
```
> fit <- glm(killsurv~conc, family="binomial")
```

Summary Output

The result for the insecticide example is:

```
> summary(glm(killsurv ~ conc, family = "binomial")

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.8923     0.6426  -7.613 2.67e-14 ***
conc          3.1088     0.3879   8.015 1.11e-15 ***
---
Null deviance: 96.6881  on 4  degrees of freedom
Residual deviance:  1.4542  on 3  degrees of freedom
AIC: 24.675
```

Proportion of Killed Insects



Global Tests for Binomial Regression

For GLMs there are three tests that can be done:

- **Goodness-of-fit test**
  - based on comparing against the saturated model
  - not suitable for non-grouped, binary data
- **Comparing two hierarchical models**
  - likelihood ratio test leads to deviance differences
  - test statistics has an asymptotic Chi-Square distribution
- **Global test**
  - comparing versus an empty model with only an intercept
  - this is a nested model, take the null deviance

Goodness-of-Fit Test

→ **the residual deviance will be our goodness-of-fit measure!**

Paradigm: take twice the difference between the log-likelihood for our current model and the saturated one, which fits the proportions perfectly, i.e. $\hat{p}_i = y_i / n_i$

$$D(y, \hat{p}) = 2\sum_{i=1}^{k}\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{(n_i - y_i)}{(n_i - \hat{y}_i)}\right)\right]$$

Because the saturated model fits as well as any model can fit, the deviance measures how close our model comes to perfection.

Evaluation of the Test

Asymptotics:

If Y_i is truly binomial and the n_i are large, the residual deviance is approximately χ^2 distributed. The degrees of freedom is:

$$k - (\#\ of\ \ predictors) - 1$$

```
> 1 - pchisq(deviance(fit), df.residual(fit))
[1] 0.69287
```

Quick and dirty:

$Deviance \gg df$: → model is not worth much.
More exactly: check $df \pm 2\sqrt{df}$

→ only apply this test if at least all $n_i \geq 5$

Overdispersion

What if $Deviance \gg df$ **???**

1) Check the structural form of the model

- model diagnostics
- predictor transformations, interactions, …

2) Outliers

- should be apparent from the diagnostic plots

3) IID assumption for p_i **within a group**

- unrecorded predictors or inhomogeneous population
- subjects influence other subjects under study

Overdispersion: a Remedy

We can deal with overdispersion by estimating:

$$\hat{\phi} = \frac{X^2}{n-p} = \frac{1}{n-p} \cdot \sum_{i=1}^{n} \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1-\hat{p}_i)}$$

This is the sum of squared Pearson residuals divided with the df

Implications:

- regression coefficients remain unchanged
- standard errors will be different: inference!
- need to use an F-test for comparing nested models

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Results when Correcting Overdispersion

```
> phi <- sum(resid(fit)^2)/df.residual(fit)
> phi
[1] 0.4847485
> summary(fit, dispersion=phi)
           Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.8923     0.4474  -10.94   <2e-16 ***
conc          3.1088     0.2701   11.51   <2e-16 ***
---
(Dispersion parameter taken to be 0.4847485)
Null deviance: 96.6881  on 4  degrees of freedom
Residual deviance:  1.4542  on 3  degrees of freedom
AIC: 24.675
```

Global Tests for Binomial Regression

For GLMs there are three tests that can be done:

- **Goodness-of-fit test**
  - based on comparing against the saturated model
  - not suitable for non-grouped, binary data
- **Comparing two hierarchical models**
  - likelihood ratio test leads to deviance differences
  - test statistics has an asymptotic Chi-Square distribution
- **Global test**
  - comparing versus an empty model with only an intercept
  - this is a nested model, take the null deviance

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Testing Hierarchical Models / Global Test

For binomial regression, these two tests are conceptually equal to the ones we already discussed in binary logistic regression.

→ *We refer to our discussion there and do not go into further detail here at this place!*

Null hypothesis and test statistic:

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

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

Distribution of the test statistic:

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

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 2014 – Generalized Linear Modeling

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!

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

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).

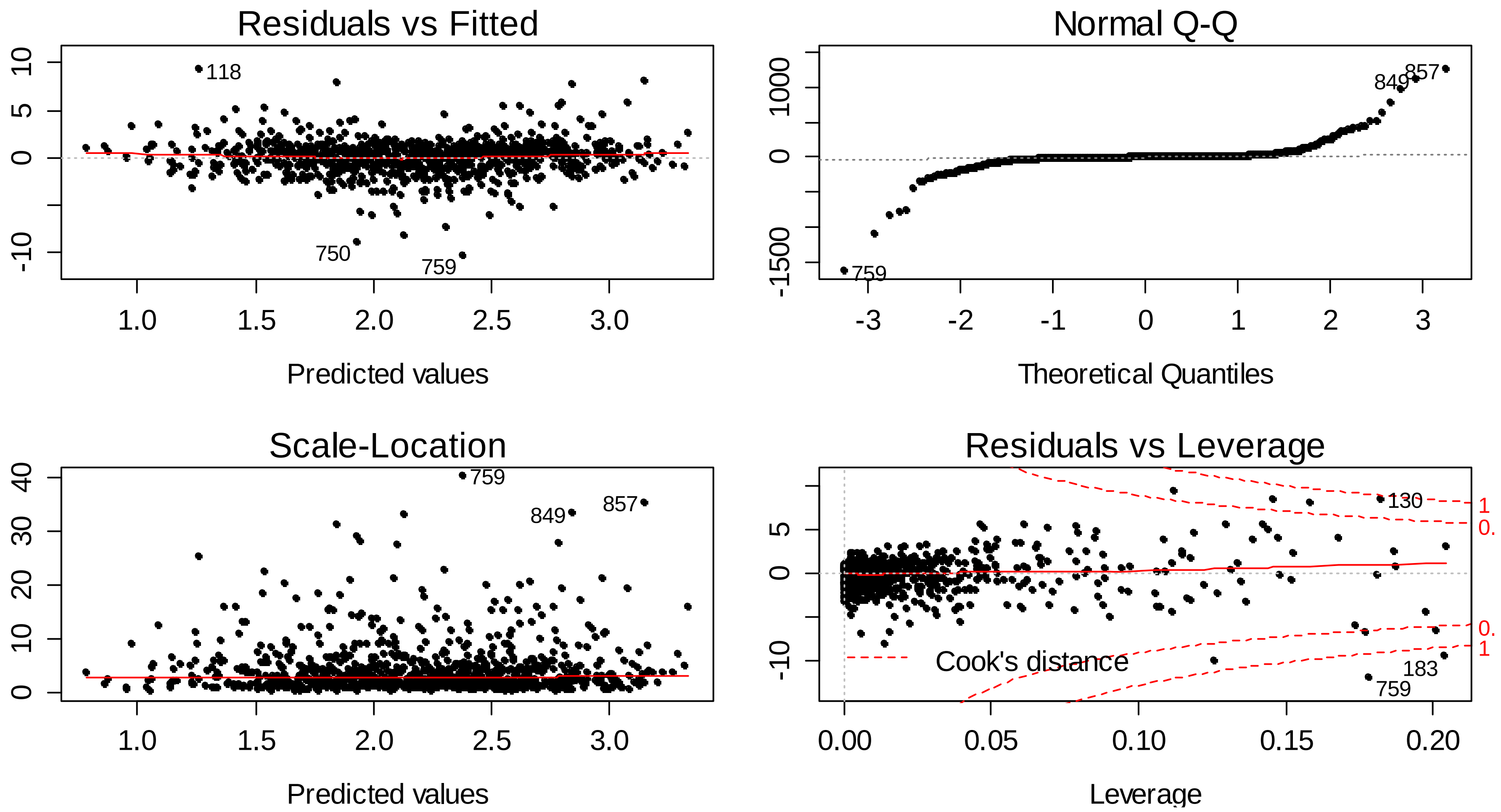
Summary Output

```
> 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!

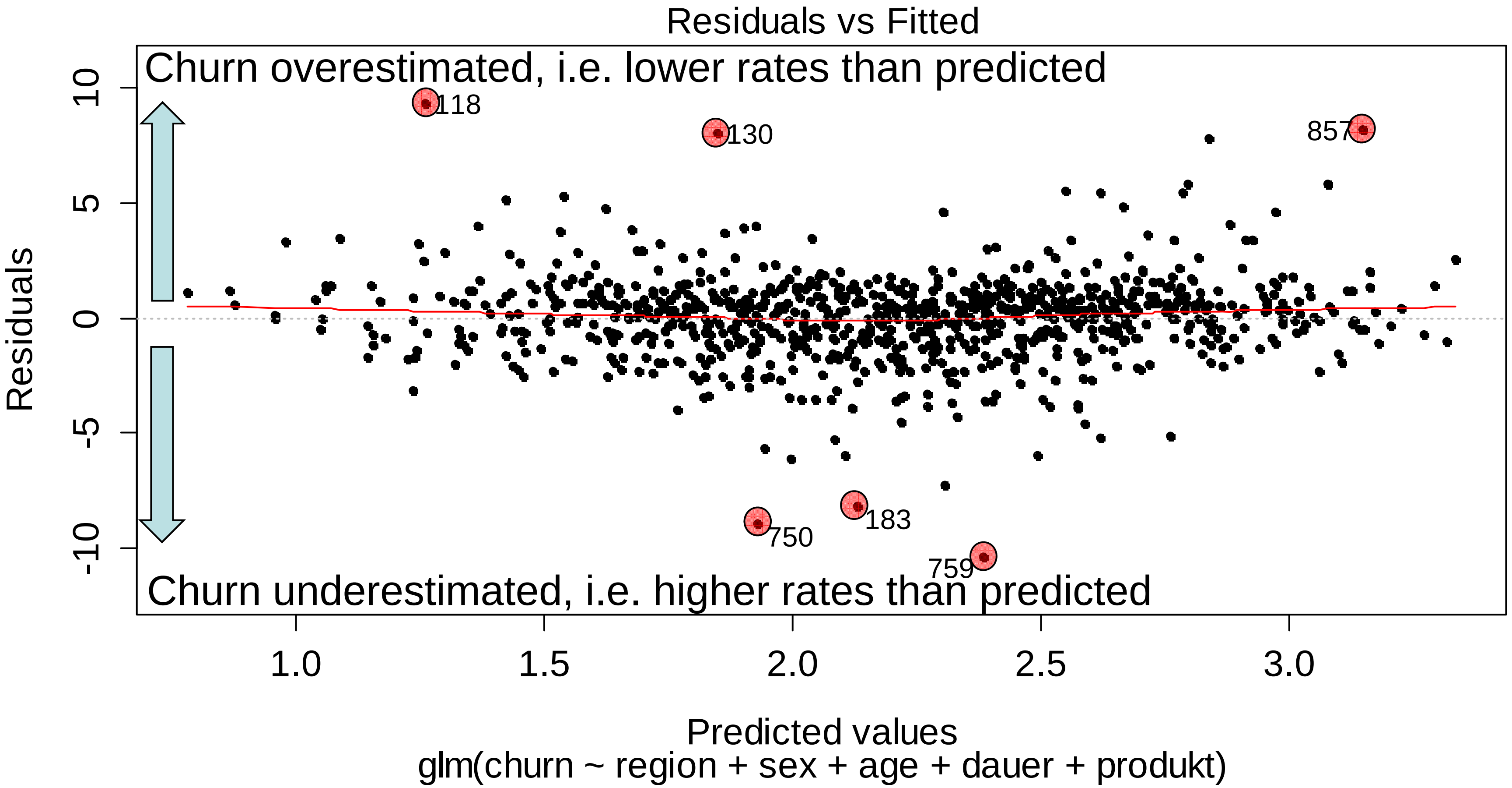
Applied Statistical Regression
AS 2014 – Generalized Linear Modeling

Model Diagnostics



Applied Statistical Regression
AS 2014 – Generalized Linear Modeling

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.

Poisson-Regression

When to apply?

- *Generally, if the* ***response variable is a count****. However:*
  - for bounded counts, the binomial model can be useful
  - for large numbers the normal approximation can serve
- *The use of Poisson regression is a must if:*
  - the counts are small and/or population size unknown
  - the population size is big and hard to come by, and the probability of an event, resp. the counts are small.

Model, Estimation, Inference:

Poisson Regression fits within the GLM framework!

Example: Tortoise Species on Galapagos

The data are as follows:

```
> library(faraway); data(gala); head(gala[,-2])
             Species  Area Elevation Nearest Scruz Adjacent
Baltra            58 25.09       346     0.6   0.6     1.84
Bartolome         31  1.24       109     0.6  26.3   572.33
Caldwell           3  0.21       114     2.8  58.7     0.78
Champion          25  0.10        46     1.9  47.4     0.18
Coamano            2  0.05        77     1.9   1.9   903.82
Daphne.Major      18  0.34       119     8.0   8.0     1.84
```

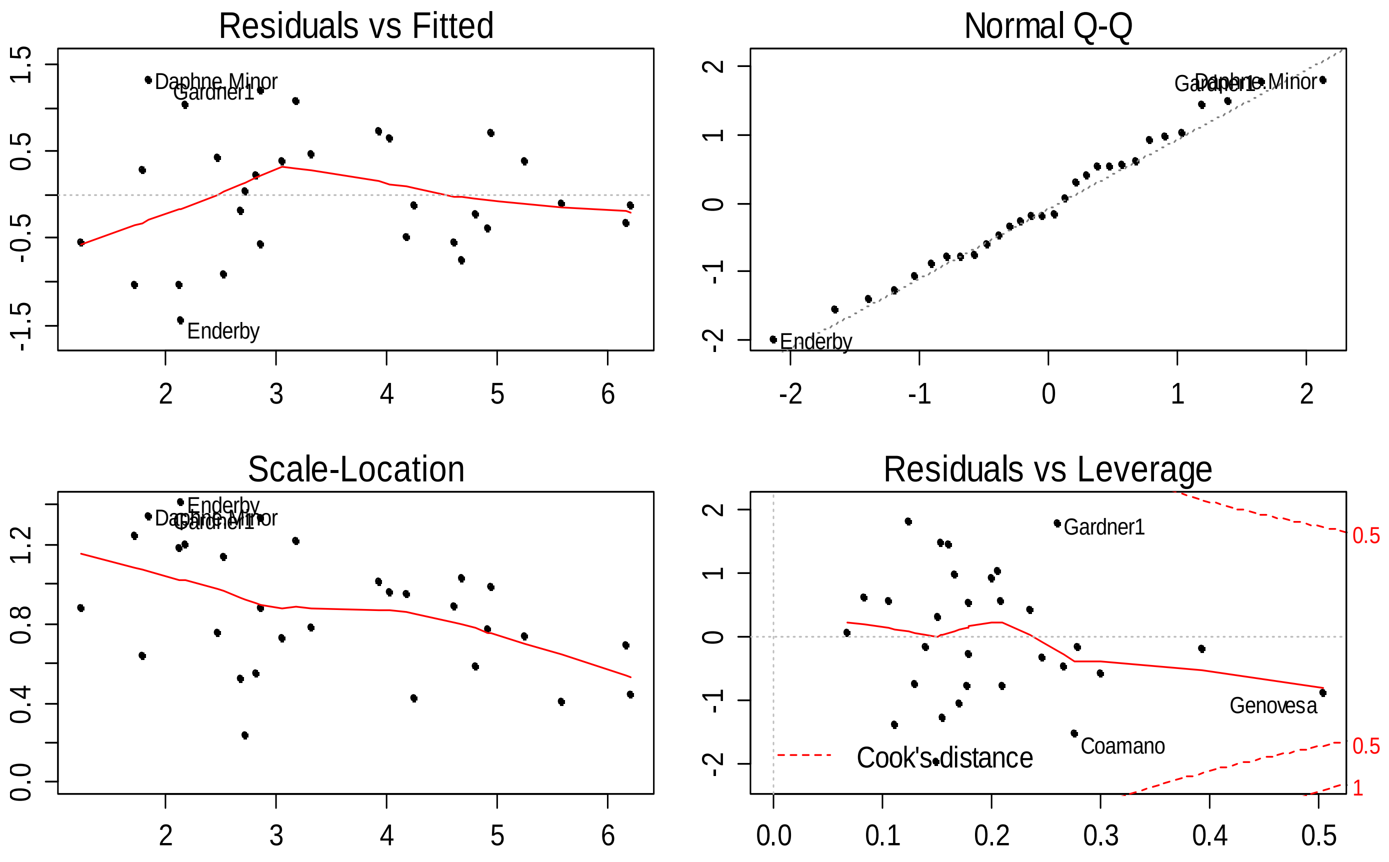
Because the predictors all take positive values only and are skewed to the right, we urgently need transformations, namely:

→ a log-transformation for all variables!

Applied Statistical Regression
AS 2014 – Generalized Linear Modeling

Trying Multiple Linear Regression

```
> fit03 <- lm(log(Species) ~ log(Area) + ..., data=gala[,-2])
```



The normal plot is fine and there are no outliers.

But it seems that the relation has a bias. The variance is ↘ .

Model needs to be improved!

Poisson Regression: Theory

We have count response: $Y_i \mid X \sim Pois(\lambda_i)$

→ The goal is to relate the parameter λ_i, which is also the conditional expectation $\lambda_i = E[Y_i \mid X]$ linearly to the predictors. Since it takes positive values only, we require a log-trsf:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

This is a GLM. The coefficients can be estimated by MLE. Assuming independence, the likelihood function is:

$$P(Y_1 = y_1, \ldots, Y_n = y_n \mid X) = \prod_{i=1}^{n} P(Y_i = y_i \mid X) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} \cdot e^{-\lambda_i}}{y_i\,!}$$

Poisson Regression: R Commands

```
> fit <- glm(Species ~ log(Area)+..., family=poisson, data=...)
> summary(fit)
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          3.323245   0.286430  11.602  < 2e-16 ***
log(Area)            0.350370   0.018005  19.459  < 2e-16 ***
log(Elevation)       0.033108   0.057034   0.580  0.56158
log(Nearest)        -0.040153   0.014071  -2.854  0.00432 **
I(log(Scruz + 0.4)) -0.035848   0.013207  -2.714  0.00664 **
log(Adjacent)       -0.089452   0.006944 -12.882  < 2e-16 ***
---
    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  359.94  on 24  degrees of freedom
AIC: 532.77
```

→ These results are based on numerical optimization.
Thus, watch the convergence of the IRLS algorithm.

Applied Statistical Regression
AS 2014 – Generalized Linear Modeling

Does the Model Fit?

Quick check: $residual\ deviance \gg df$???

More precisely: $D = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right] \sim \chi^2_{n-(p+1)}$

Thus, when testing H_0 :"*Model is correct*", we obtain:

```
> pchisq(359.94, 24, lower=FALSE)
[1] 1.185031e-61
```

→ The ***model does not fit well***. There is (much) more variation in the response than the Poisson distribution alone suggests. *Why is this and where does it come from?*

→ Diagnostic plots / visualization is key!

Residual Analysis

Analyze deviance or (as in R) Pearson residuals:

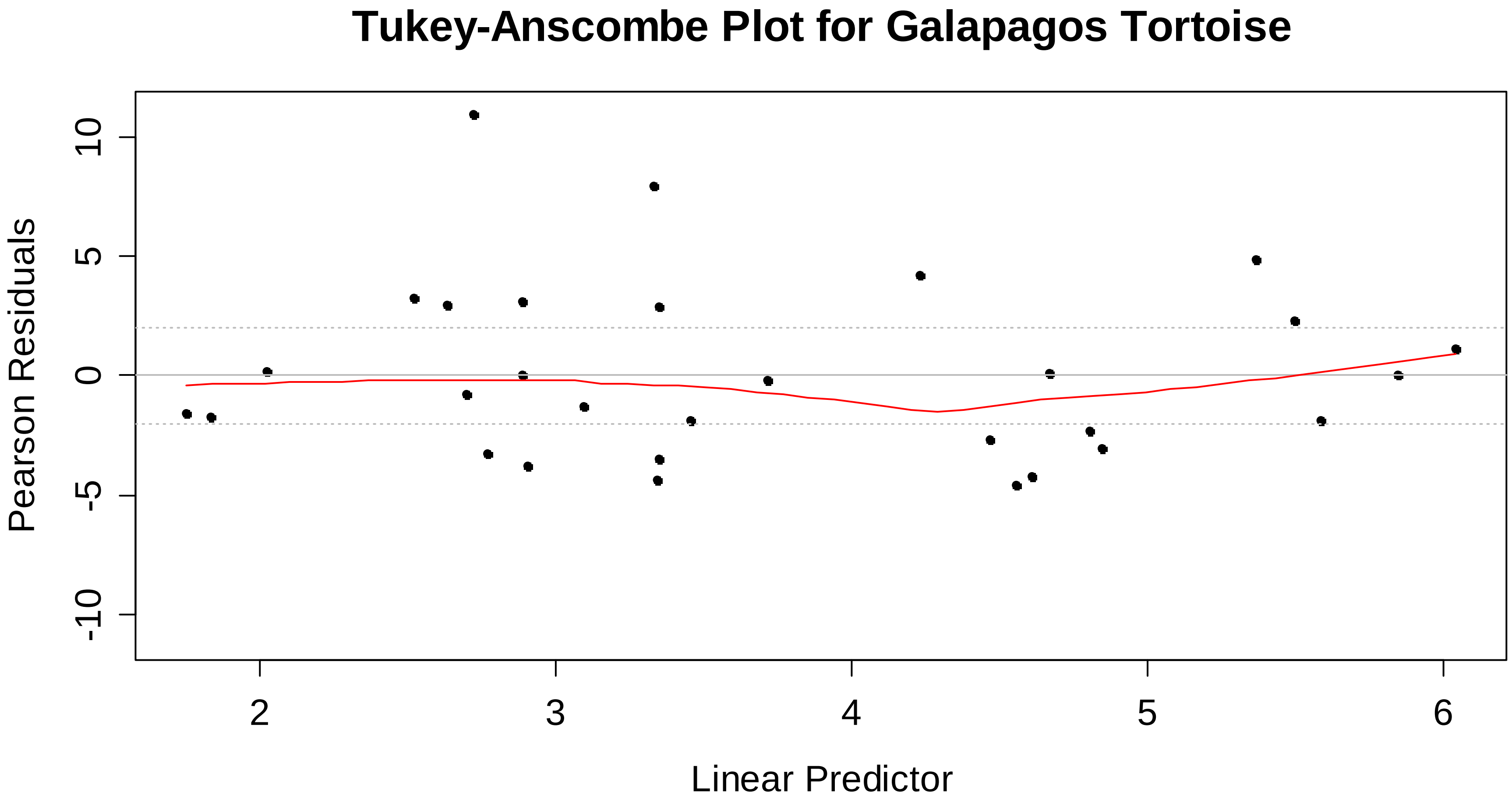
$$P_i = \frac{(y_i - \hat{\lambda}_i)}{\sqrt{\hat{\lambda}_i}} \quad \text{approx.} \sim N(0,1)$$

Thus, residuals $|P_i| > 2$ are bigger than the Poisson distribution suggests. And even larger residuals $|P_i| > 4$ would not exists if the Poisson model was correct.

```
> xx <- predict(fit, type="link")
> yy <- resid(fit, type="pearson")
> plot(xx, yy, main="Tukey-Anscombe Plot...")
> lines(loess.smooth(xx, yy), col="red")
```

Applied Statistical Regression
AS 2014 – Generalized Linear Modeling

Tukey-Anscombe Plot



Dealing with Overdispersion

If the predictor-response relation is correct, but the variation is observed to be bigger than the distribution model suggests:

$\hat{\beta}_0,...,\hat{\beta}_p$ and $\hat{\lambda}_i$ unbiased

Standard errors $se(\hat{\beta}_0),...,se(\hat{\beta}_p)$ are wrong

Standard errors are corrected using a dispersion parameter:

$$\hat{\phi} = \frac{\sum_i (y_i - \hat{\lambda}_i)^2 / \hat{\lambda}_i}{n-(p+1)}$$

In R:

```
> sum(resid(fit, type="pearson")^2)/fit$df.res
[1] 16.64651
```

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Final Result

```
> summary(fit, dispersion=16.64651)
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          3.32325    1.16864   2.844  0.00446 **
log(Area)            0.35037    0.07346   4.769 1.85e-06 ***
log(Elevation)       0.03311    0.23270   0.142  0.88686
log(Nearest)        -0.04015    0.05741  -0.699  0.48430
I(log(Scruz + 0.4)) -0.03585    0.05389  -0.665  0.50589
log(Adjacent)       -0.08945    0.02833  -3.157  0.00159 **
---
Dispersion parameter for poisson family: 16.647
Null deviance: 3510.73 on 29 degrees of freedom
Residual deviance: 359.94 on 24 degrees of freedom
AIC: 532.77
```

→ Inference result is similar to the one from multiple linear regression. **Mathematics says: this is not a surprise!**

Informations on the Exam

- The exam will be on Saturday (!!!) January 24, 2015 at 9am. Please do not rely on this, but see the official announcement.
- It will be open book, i.e. you are allowed to bring any written materials you wish. You can also bring a pocket calculator, but computers/notebooks and communcation aids are forbidden.
- Topics include everything that was presented in the lectures, from the first to the last, and everything that was contained in the exercises and master solutions.
- You will not have to write R-code, but you should be familiar with the output and be able to read it.

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

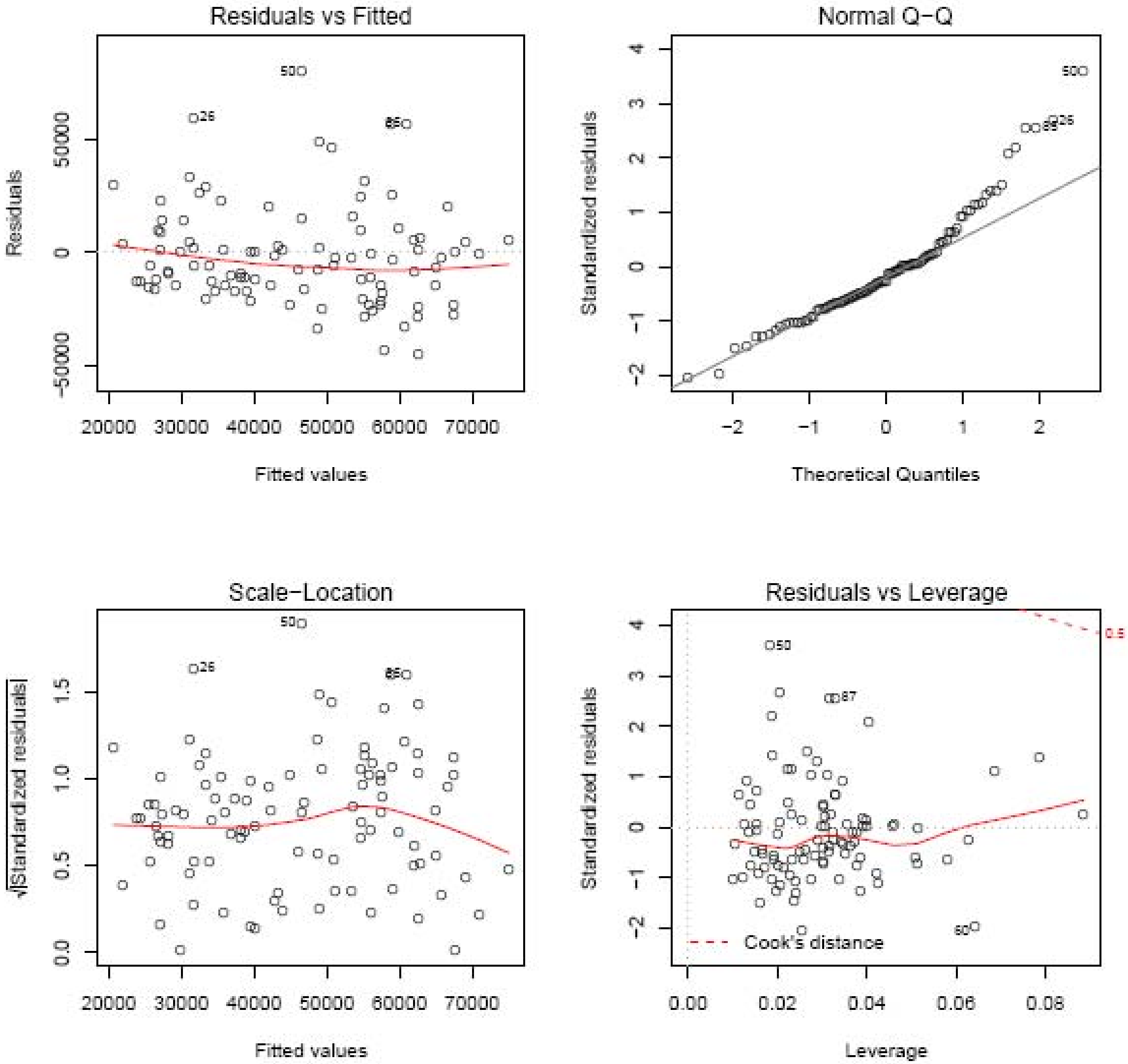
Informations on the Exam

- With the exam, we will try our best to check whether you are proficient in applied regression. This means choosing the right models, interpreting output and suggesting analysis strategies.
- Some old exams are available for preparation. I recommend that you also make sure that you understand the lecture examples well and especially focus on the exercises.
- Concerning the old exams, please keep in mind that both contents and style of the course/questions may evolve.
- There will be question hours in January. See the course webpage where time and location will be announced.

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Sample Questions from Previous Exams



Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Sample Questions from Previous Exams

Looking at the plots: Which of the statements are correct?

a) The normality assumption of the errors is heavily violated.

b) The errors are not independent.

c) The assumption of constant error variance is heavily violated.

d) There are clear outliers.

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Sample Questions from Previous Exams

You would like to make predictions with your model. Would you do anything beforehand in order for the model assumptions to be better fulfilled?

a) A transformation of the response seems to be reasonable as a first action.

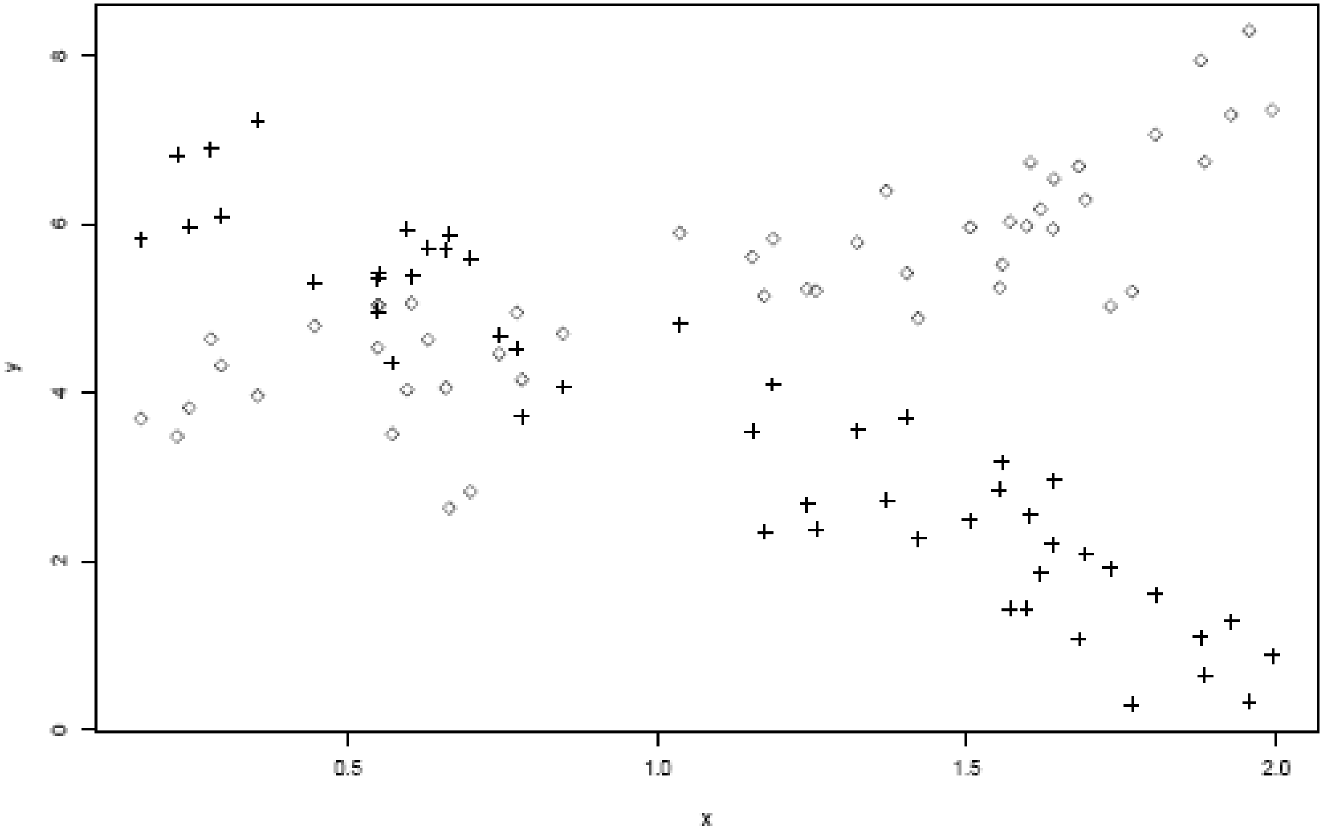
b) If one is only interested in predictions, the model assumptions are not important. These are only important for tests.

c) Because no leverage points are detectable in the leverage-plot, the model is not changing much if actions are taken to better full the model assumptions.

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Sample Questions from Previous Exams



Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Sample Questions from Previous Exams

The different symbols in the plot correspond to the values of the different groups.

a) What model would you fit to these data?

b) What is the model equation?

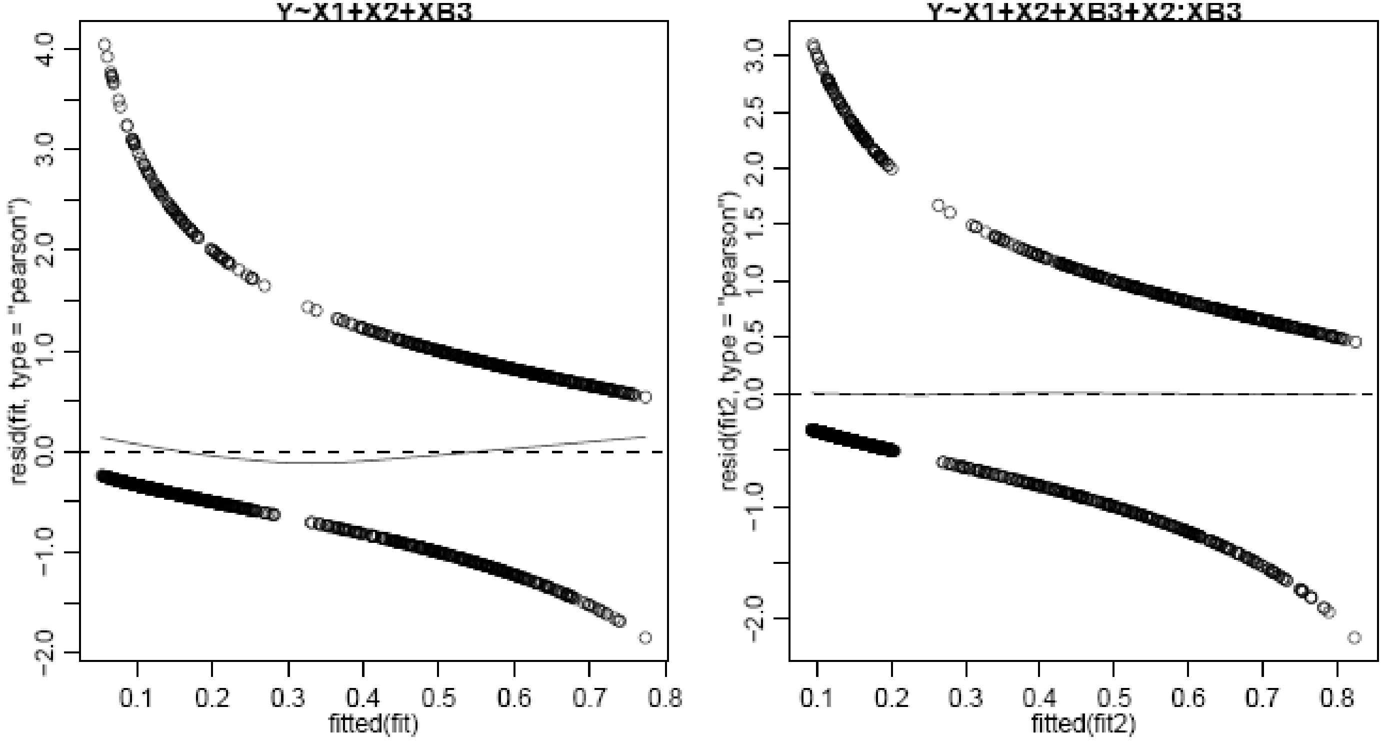
c) Which regression coefficients in your model are clearly positive, cleary negative, approximately 0?

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

Sample Questions from Previous Exams

e) Which of the two models do you prefer and why? Decide based on the output and the following plots of the Pearson residues versus $\hat{p}_i$.



f) In the first model X_2 is significant, but in the second model it is not. Interpret why (one to two sentences)!

Applied Statistical Regression

AS 2014 – Generalized Linear Modeling

End of the Course

→ **Happy holidays and all the best for the exams!**