Applied Statistical Regression
HS 2011 – 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 19, 2011

Applied Statistical Regression
HS 2011 – Week 13

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: there is more than 1 observation for a given predictor setting

Applied Statistical Regression
HS 2011 – Week 13

Model and Estimation

The goal is to find a relation:

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

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:

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

Applied Statistical Regression
HS 2011 – Week 13

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

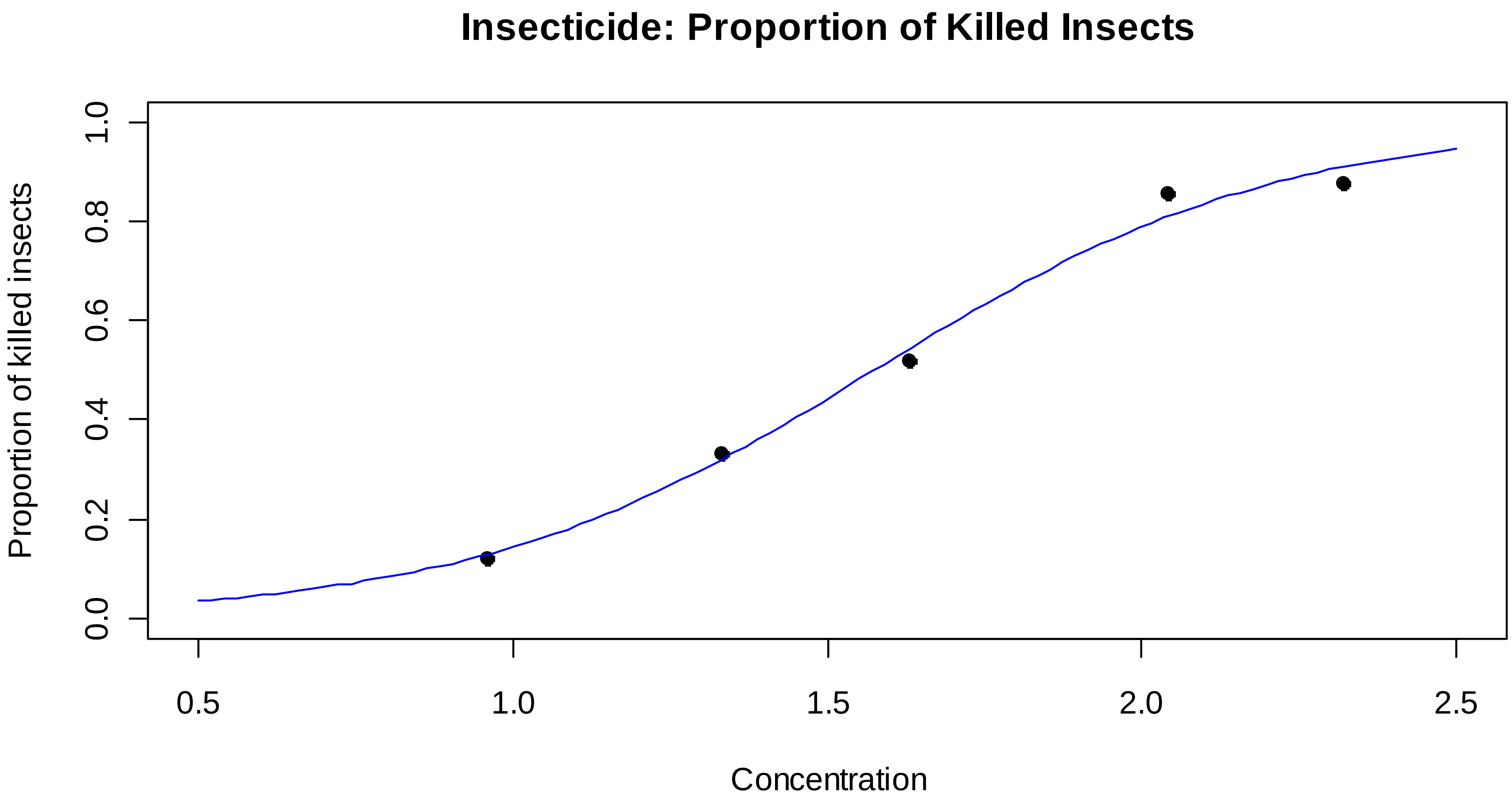
The result for the insecticide example is:

```
> 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
```

Applied Statistical Regression
HS 2011 – Week 13

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 nested 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 deviance is approximately χ^2 distributed. The degrees of freedom is:

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

```
> pchisq(deviance(fit), df.residual(fit), lower=FALSE)
[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 a test for comparing nested models

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 nested 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

Testing Nested Models and the 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:

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

Poisson-Regression

When to apply?

- Responses need to be counts
  - 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:
  - unknown population size and small counts
  - when the size of the population is large and hard to come by, and the probability of "success"/ the counts are small.

Methods:

Very similar to Binomial regression!

Extending...: Example 2

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 linear way on the covariates:

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

Applied Statistical Regression
HS 2011 – Week 13

Informations on the Exam

- The exam will be on February 7, 2012 (provisional) and lasts for 120 minutes. But please 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
HS 2011 – Week 13

Informations on the Exam

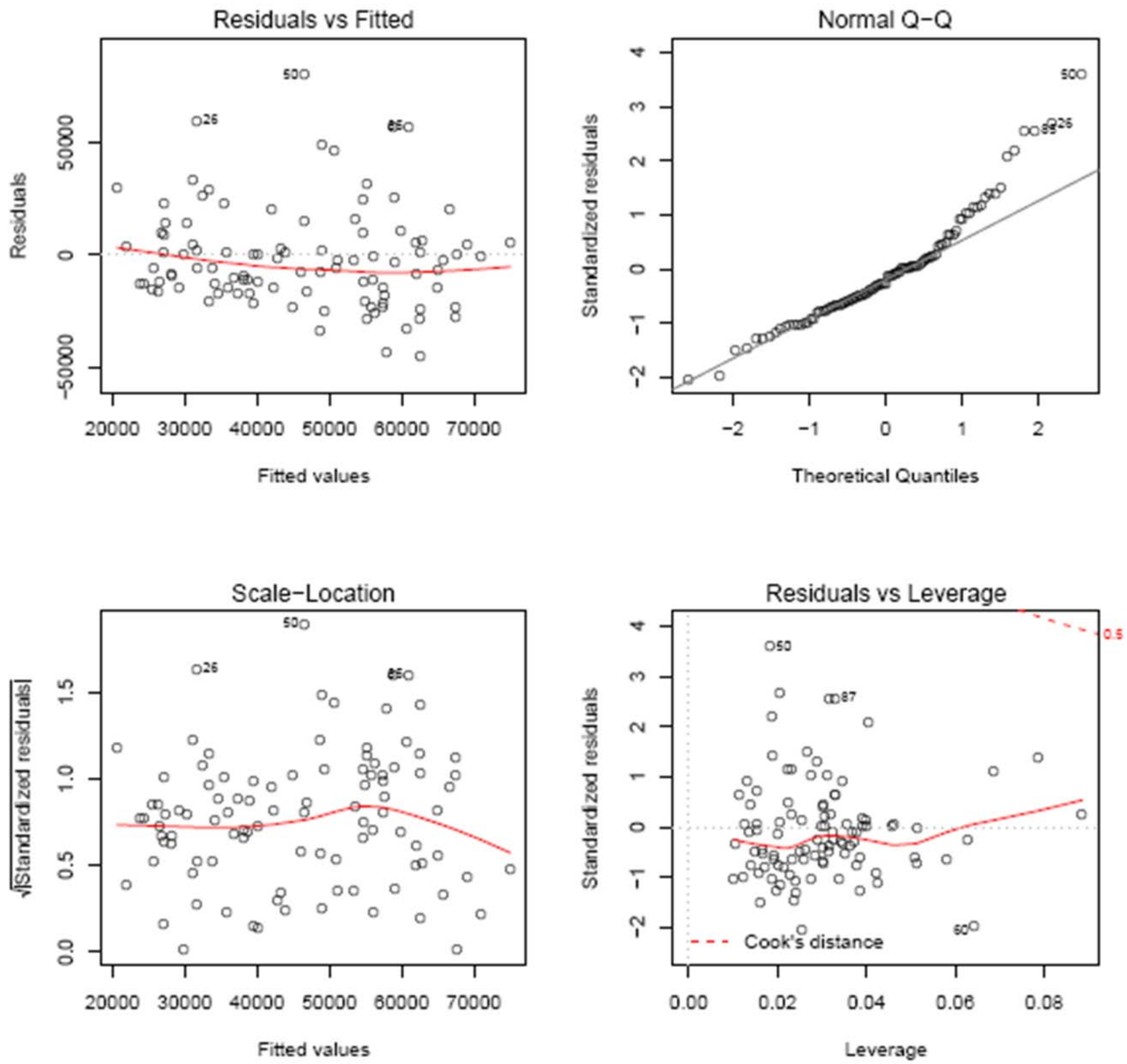
- With the exam, we will try our best to check whether you are proficient in applied regression. This means choosing the right models, interpreting output and suggesting analysis strategies.

- Two old exams will be available for preparation. I recommend that you also make sure that you understand the lecture examples well and especially focus on the exercises.

- There will be question hours in January. See the course webpage where time and location will be announced.

Applied Statistical Regression
HS 2011 – Week 13

Sample Questions from Previous Exams



Applied Statistical Regression
HS 2011 – Week 13

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
HS 2011 – Week 13

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 fullfilled?

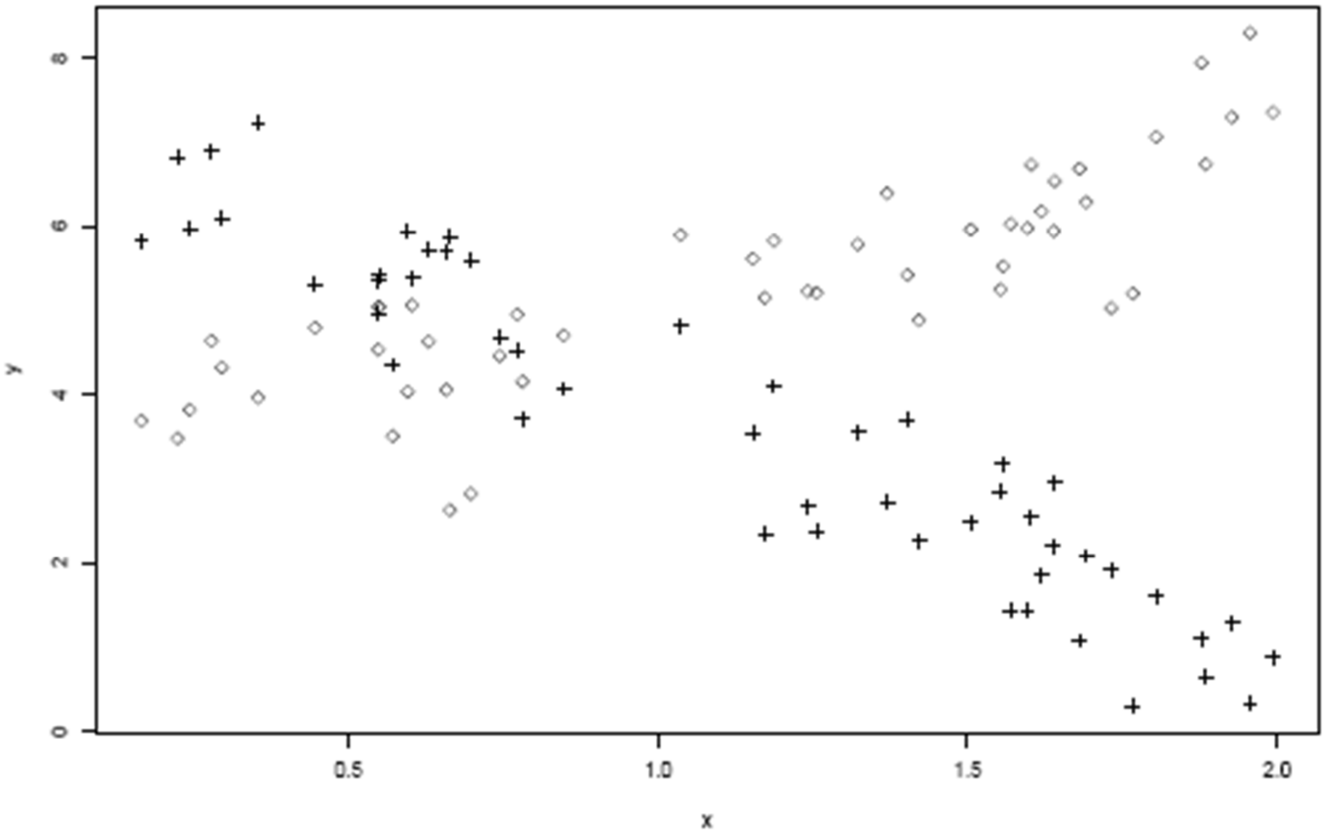
a) A transformation of the response seems to be reasonable as a first action.

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
HS 2011 – Week 13

Sample Questions from Previous Exams



Sample Questions from Previous Exams

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

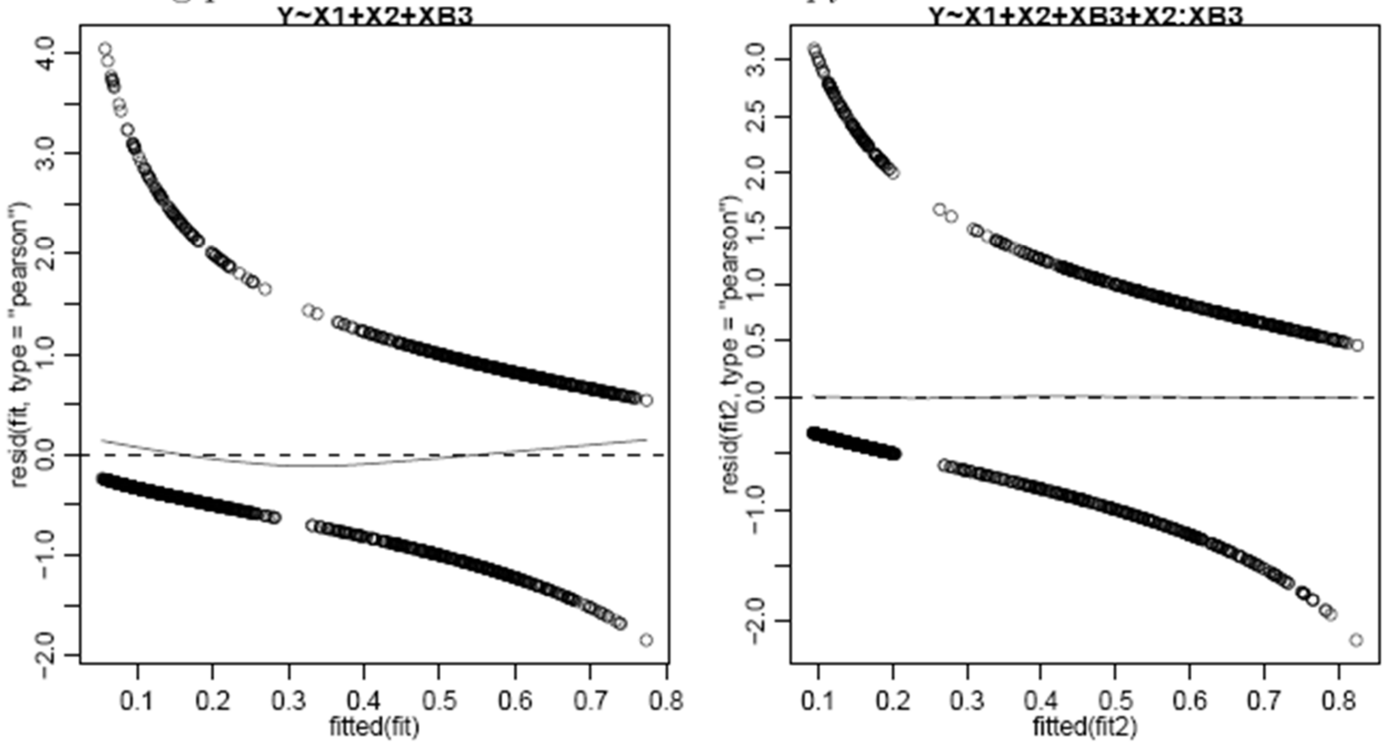
a) What model would you fit to these data?

b) What is the model equation?

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

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
HS 2011 – Week 13

End of the Course

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