Applied Statistical Regression
AS 2012 – 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 17, 2012

Applied Statistical Regression

AS 2012 – Week 13

Poisson-Regression

When to apply?

- *The response variable is a count*
  - 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!

Applied Statistical Regression

AS 2012 – Week 13

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 to take FATs, namely:

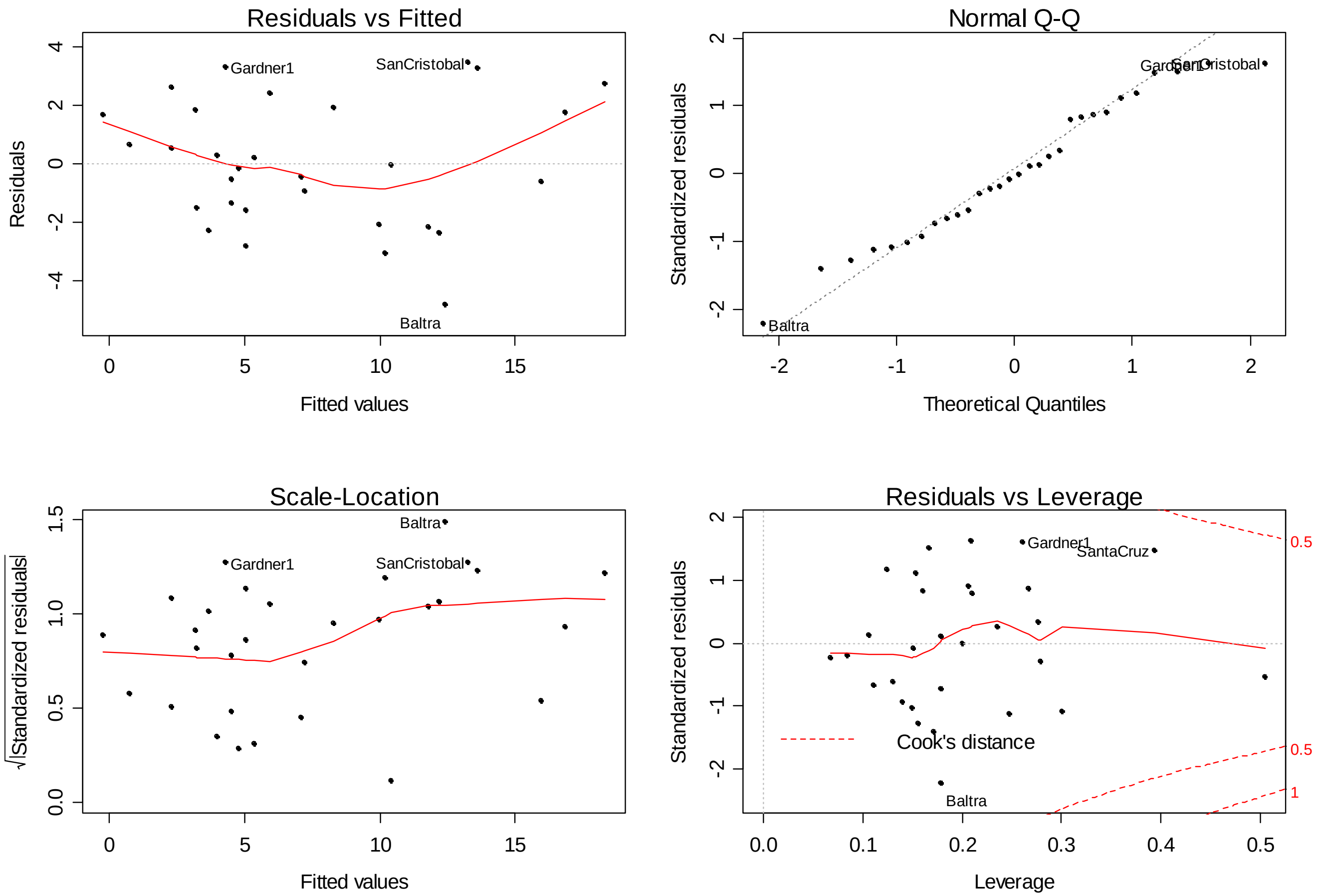
→ see blackboard...

Applied Statistical Regression

AS 2012 – Week 13

Trying Multiple Linear Regression

```
> fit02 <- lm(sqrt(Species) ~ log(Area) + ..., data=gala[,-2])
```



Variance and distribution of the errors are +/- OK.

There is a pronounced bias in the relation.

Model needs to be improved!

Applied Statistical Regression
AS 2012 – Week 13

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 2012 – Week 13

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!

Applied Statistical Regression
AS 2012 – Week 13

Residual Analysis

Analyze deviance or (as in R) Pearson residuals:

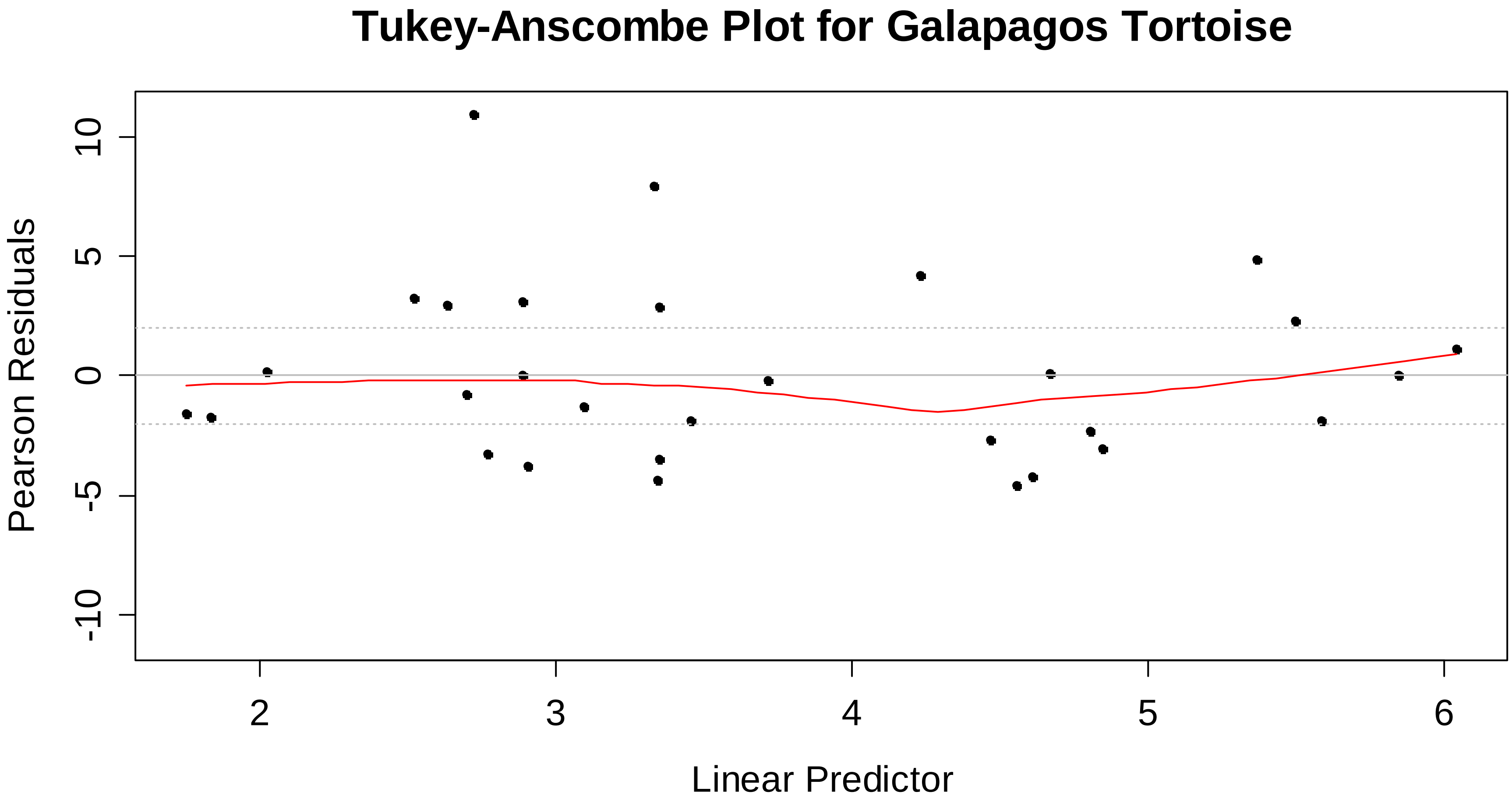
$$P_i = \frac{(y_i - \hat{\lambda}_i)}{\sqrt{\hat{\lambda}_i}} \quad \text{approx.} \sim N(0,1)$$

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 2012 – Week 13

Tukey-Anscombe Plot



Applied Statistical Regression

AS 2012 – Week 13

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 2012 – Week 13

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!**

Analysis of Multinomial Response

- Response $Y_i \in \{1, ..., J\}$ is categorical with more than 2 levels.
- **Nominal multinomial data:**

  → response does not have a natural ordering
  e.g. *car makes*, *colors*, *political party*…
- **Ordinal multinomial data:**

  → response categories can be ordered
  e.g. *avalanche danger*
- Can be seen as an extension to logistic/binomial regression

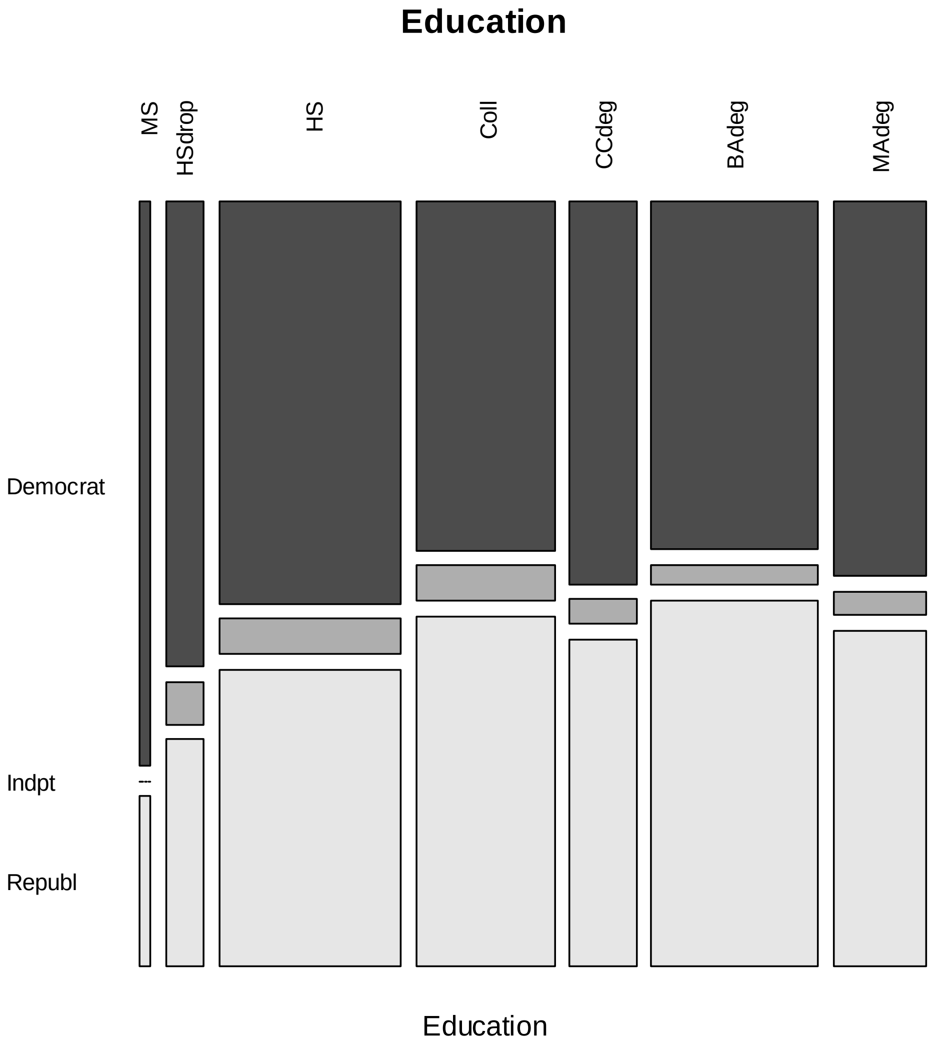
Example

American National Election Study 1996: 944 observations

- Response variable: ***party identification***
  - Democrat / Independent / Republican
- Predictor 1: ***education***
  - 7 levels: middle school – high school drop - … - MA degree
- Predictor 2: ***income***
  - pseudo-continuous with 24 different values, yearly income
- Predictor 3: ***age***
  - continuous, age in years

Applied Statistical Regression
AS 2012 – Week 13

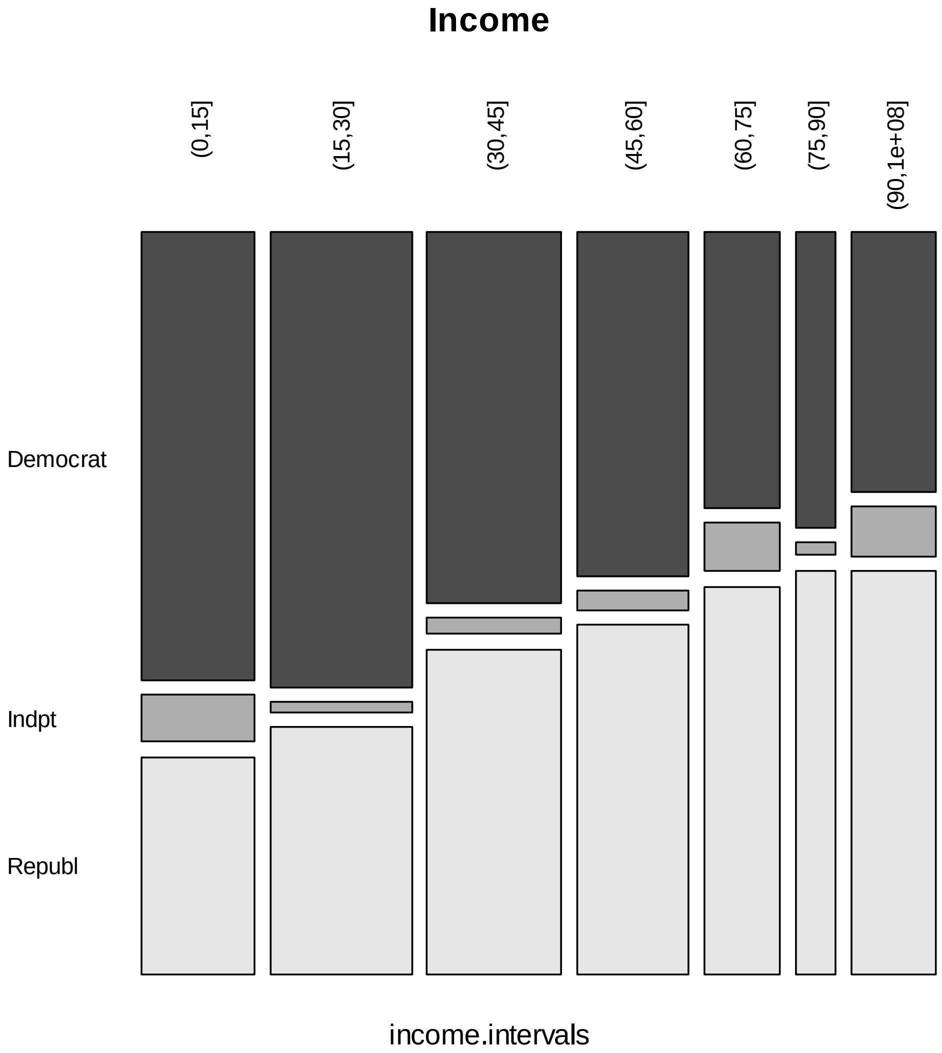
Mosaic Plot of Education



Applied Statistical Regression

AS 2012 – Week 13

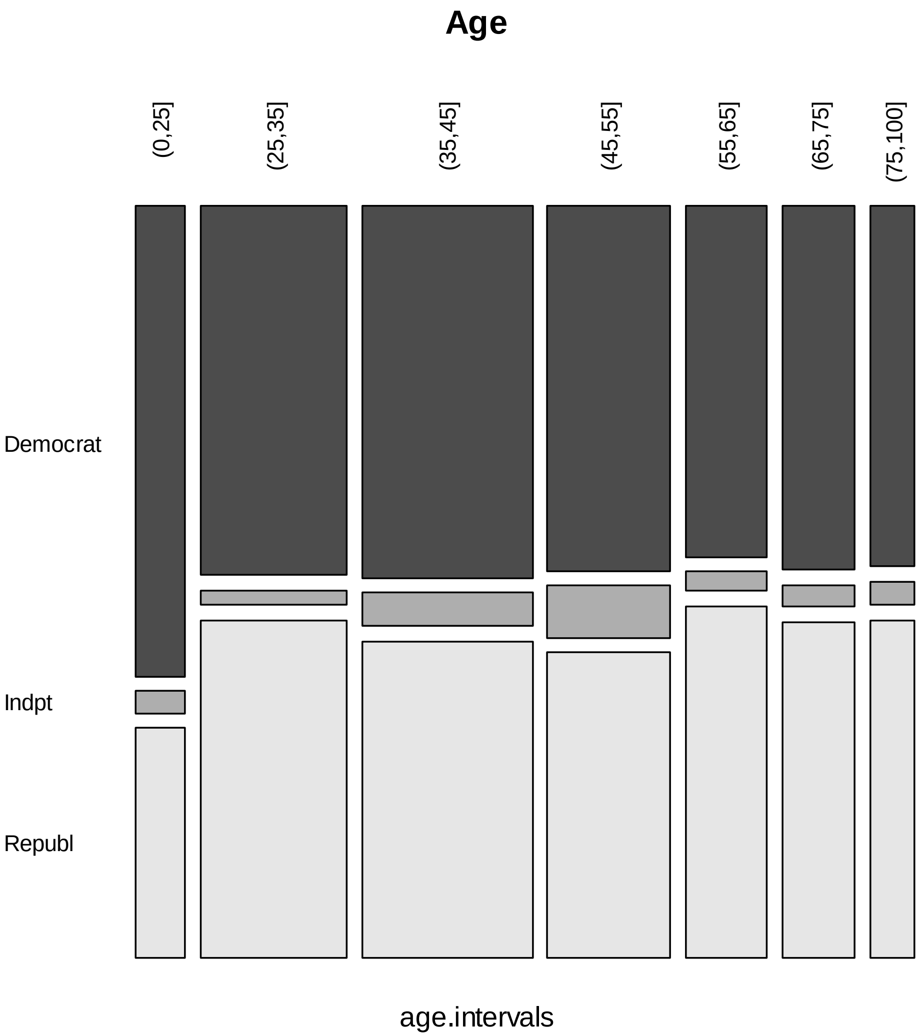
Mosaic Plot of Income



Applied Statistical Regression

AS 2012 – Week 13

Mosaic Plot of Age



Cross-Sectional vs. Longitudinal Data

Cross-sectional data:

We observe persons of different age/income and ask their party identification, but only once in their lifetime.

Longitudinal data:

We observe persons some persons over a long time period and determine how age, income & party identification change.

What can we say?

We cannot say anything about what will happen with an individual when it gets older or develops to a higher income, but can only give the relative probability of party affilitation.

Applied Statistical Regression

AS 2012 – Week 13

Multinomial Logit Model

- Response $Y_i \in \{1, ..., J\}$
- Ultimate goal: probabilities $p_{ij} = P(Y_i = j \mid X)$
- There can be grouped and non-grouped data
- Y_{ij} is the number of observations in category j for group/ind. i
- $n_i = \sum_j Y_{ij}$ is the number of individuals in group i

The Y_{ij}, conditional on the n_i, have a multinomial distribution:

$$P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, ..., Y_{iJ} = y_{iJ}) = \frac{n_i}{y_{i1}! \cdot ... \cdot y_{iJ}!} p_{i1}^{y_{i1}} \cdot ... \cdot p_{iJ}^{y_{iJ}}$$

Applied Statistical Regression

AS 2012 – Week 13

Using the Logit Transformation

As with binomial data, our goal will again be to find a relation between the probabilities p_{ij} and the predictors x_i, while ensuring that the probabilities are restricted to values between 0 and 1.

$$\log\left(\frac{P(Y_i = j \mid X)}{P(Y_i = 1 \mid X)}\right) = \log\left(\frac{p_{ij}}{p_{i1}}\right) = \eta_{ij} = \beta_{0j} + \beta_{1j}x_{i1} + \ldots + \beta_{pj}x_{ip}$$

This is a logit model for probability quotients, where we compare each of the categories against the first one, which serves as the reference category. The use of such a baseline category is dictated by the constraint that $\sum_j p_{ij} = 1$.

Applied Statistical Regression

AS 2012 – Week 13

Remarks to the Model

- This is an equation system with $J-1$ rows, and different coefficients for each class j.

- Quite a few parameters are thus estimated. Their number is:

$$p^* \cdot (J-1)$$

- It is (as always) better to make sure that at least 5 observations per estimated parameter are present for model fitting

- Choice of the baseline class is free. R uses the first levels in the factor variable that contains the response variable!

Fitting the Model

```
> library(nnet)
> fit <- multinom(party ~ age + income + educ,
                                  data=nes)

# weights:  30 (18 variable)
initial  value 1037.090
iter   10 value  783.325
iter   20 value  756.095
iter   30 value  755.807
final value      755.806
converged
```

Values of the log-likelihood

Applied Statistical Regression

AS 2012 – Week 13

Summary Output

```
> summary(fit)
Coefficients:
      (Intrcpt)     age   income   educ.L   educ.Q   educ.C
Indpt    -5.136   0.005    0.016    5.244   -6.341    4.693
Republ   -1.409   0.010    0.013    0.564   -0.720    0.017

         educ^4  educ^5   educ^6
Indpt    -2.552   1.291   -0.539
Republ    0.000  -0.103   -0.129

Std. Errors: ...

Residual Deviance: 1511.612
AIC: 1547.612
```

Applied Statistical Regression

AS 2012 – Week 13

Inference

No individual hypothesis tests, although standard errors are provided in the summary output!

Reason: all parameters $\beta_{k2}, ..., \beta_{kJ}$ simultaneously need to be equal to zero, which cannot be tested with an individual hypothesis test.

Way out: resort to a comparison of nested models, which will as before be based on log-likelihood ratios, resp. deviance differences. Since such models are hierachical, we can use the χ^2 approximation for testing the null hypothesis.

Inference: Example

```
> fit.age.inc <- multinom(party ~ age + income, data=nes)
> deviance(fit.age.inc) - deviance(fit)
[1] 13.70470
> pchisq(13.70470, fit$edf - fit.age.inc$edf, lower=FALSE)
[1] 0.3199618
```

- Null hypothesis: *small model is enough, 12 coefficients* $\beta_{kj} = 0$
- The p-value is 0.32, thus, **`education`** is *not significant*
- Is this a surprise, given the mosaic plot from above?
- no, the biggest differences in party affiliation are among the young people below 25 years of age, which represent only a very small fraction of the observations

Prediction

One of the predominant goals with multinomial logit models is to obtain predicted probabilities. We here show them for some arbitrary 6 instances out of the 944 that are present in total.

```
> round(predict(fit, type="probs"),3)[sample(1:944)[1:6],]
    Democrat Indpt Republ
743    0.339 0.058  0.603
239    0.524 0.018  0.457
659    0.515 0.036  0.449
174    0.513 0.024  0.462
903    0.282 0.042  0.676
863    0.345 0.037  0.618
```

Class Prediction

When we for a person need to predict which party he/she is going to vote for, we would just choose the one with the highest probability. This is easy to obtain from R:

```
> predict(fit, type="class")[sample(1:nrow(nes))[1:10]]
 [1] Republ   Democrat   Democrat   Democrat   Republ
     Republ   Democrat   Democrat   Republ   Republ
```

Model Diagnostics

- Model diagnostics are (too) difficult and “never” done in the context of multinomial logit models
- The reason is that there is no meaningful definition of what residuals are in this context
- There are some residuals for each equation, and they also depend on the choice of the baseline category.
- **How these residuals could be displayed in comprehensive form is unclear. Thus, we here remain without effective tools for model enhancement.**

Applied Statistical Regression

AS 2012 – Week 13

Multinomial Data

- Response $Y_i \in \{1, ..., J\}$ is categorical with more than 2 levels.
- **Nominal multinomial data:**

  → response does not have a natural ordering
  e.g. car makes, colors, …
- **Ordinal multinomial data:**

  → response categories can be ordered
  e.g. avalanche danger
- These are extensions to logistic/binomial regression

Example

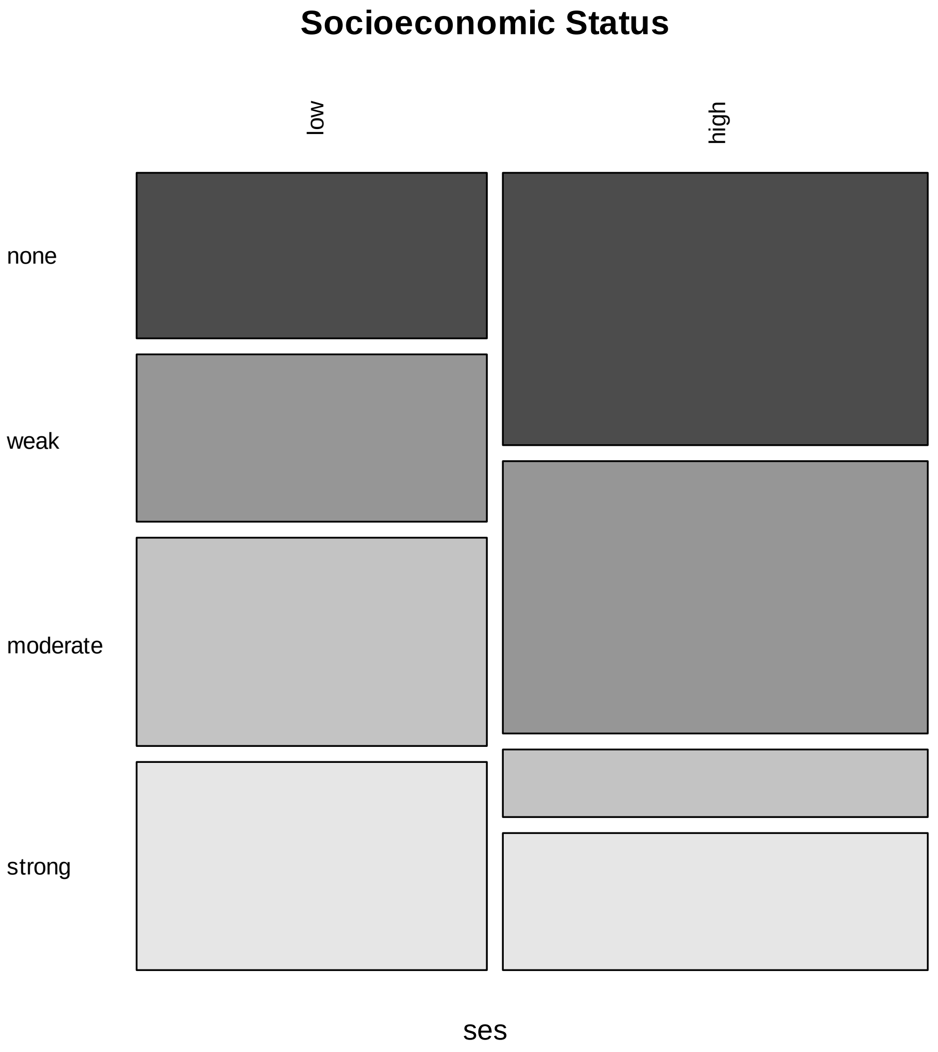
Mental Impairment Data: 40 observations

- Response variable: ***mental impairment***
  - none / weak / moderate / strong
- Predictor 1: ***socioeconomic status***
  - 2 levels: low / high
- Predictor 2: ***number of traumatic experiences in life***
  - count of potentially traumatic events such as death in family, divorce, periods of unemployment, etc.

Applied Statistical Regression

AS 2012 – Week 13

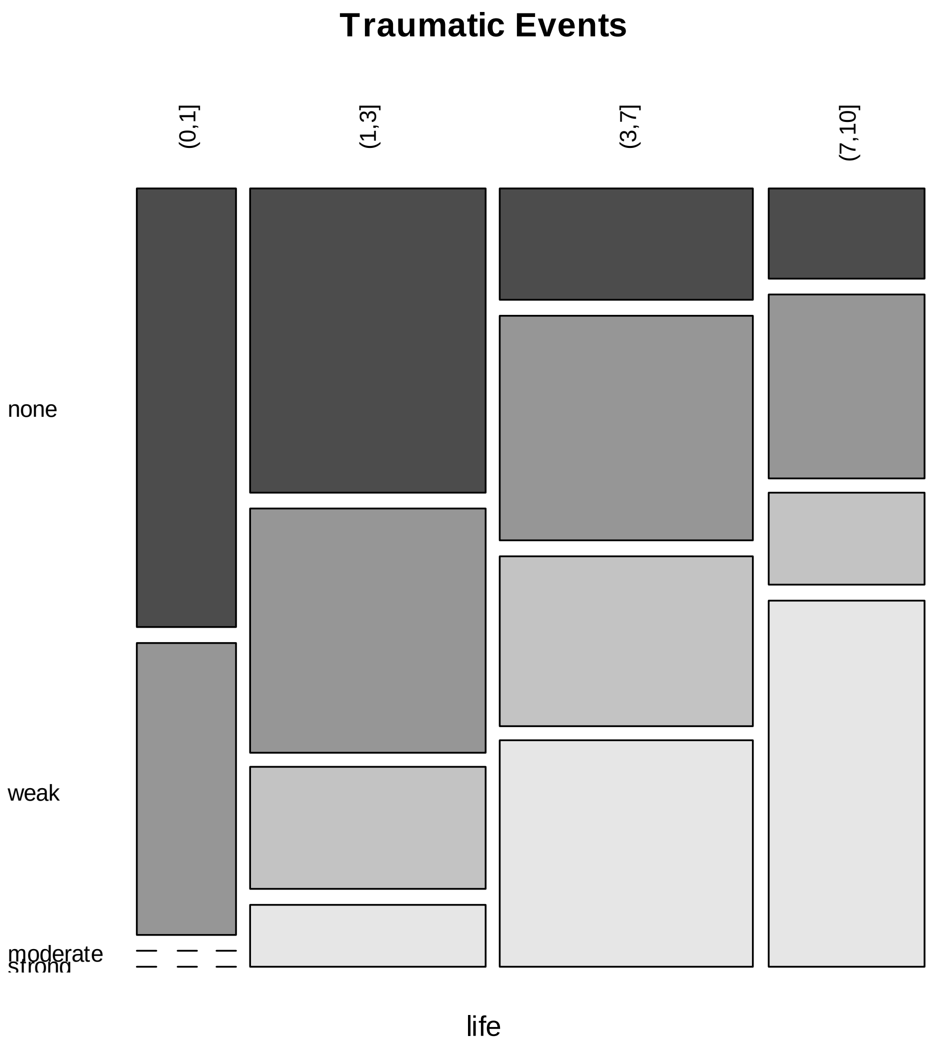
Mosaic Plot of SES



Applied Statistical Regression

AS 2012 – Week 13

Mosaic Plot of Life Events



A Model for Ordinal Responses

- Response $Y_i \in \{1, ..., J\}$: ordered categories
- Ultimate goal: probabilities $p_{ij} = P(Y_i = j)$
- With ordered response, it is easier and more powerful to work with cumulative probabilities, i.e.:

$$\gamma_{ij} = P(Y_i \leq j)$$

- The goal will be to link these cumulative probabilities to a linear combination of the predictors:

$$g(\gamma_{ij}) = \alpha_j - x_i^T \beta$$

Why this Model?

This model is much easier to comprehend if we use the notion of a latent variable Z_i. It may be thought of as the underlying continuous, but unobserved, response. In practice, we are limited to observing Y_i which are a discretized version of Z_i, and we have:

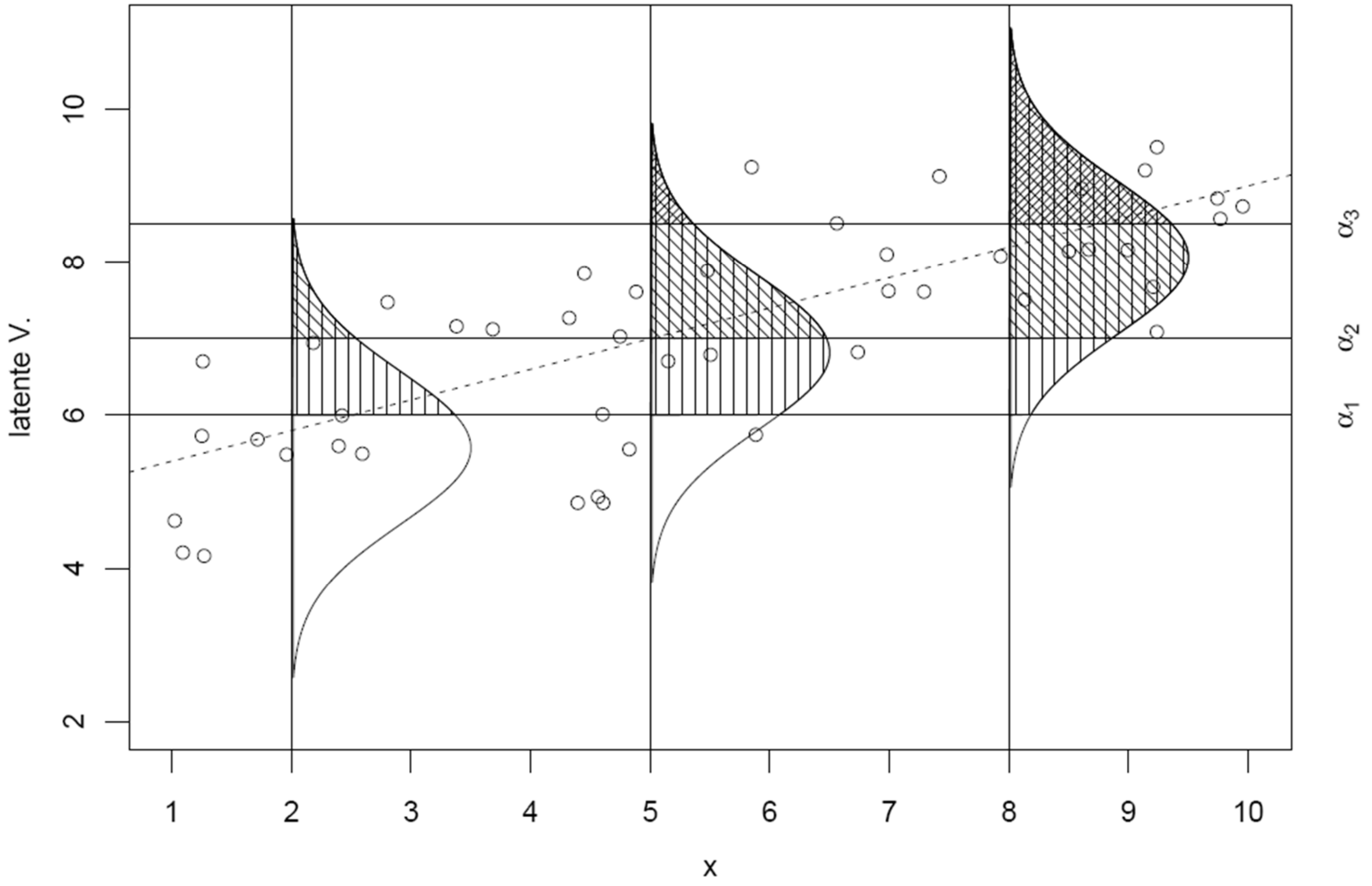
$$Y_i = j \quad \text{if} \quad \alpha_{j-1} < Z_i \le \alpha_j$$

The relation between the latent variable Z_i and the predictors is given by some multiple linear regression model, i.e.

$$Z_i = x_i^T \beta + E_i$$

Applied Statistical Regression
AS 2012 – Week 13

Latent Variable Notion



Proportional Odds Model

We are now considering the event $\{Y_i \le j\}$, which is equivalent to $\{Z_i \le \alpha_j\}$. With some algebra, we obtain:

$$\gamma_{ij} = P(Y_i \le j) == P(Z_i \le \alpha_j) = P(E_i \le \alpha_j - x_i^T \beta) = F(\alpha_j - x_i^T \beta)$$

where $F(\cdot)$ is the cumulative distribution function of the E_i.

There are 3 options:

- Logistic distribution: use the logit link function
- Gaussian distribution: use the probit link function
- Extreme value distribution: complementary log-log link

Proportional Odds Model

When we choose the logistic distribution, which has cdf:

$$F(x) = e^x / (1 + e^x),$$

we obtain the proportional odds model:

$$\gamma_{ij} = \frac{\exp(\alpha_j - x_i^T \beta)}{1 + \exp(\alpha_j - x_i^T \beta)}$$

This model can be fitted in R with function polr():

```
library(MASS)
fit <- polr(mental ~ ses + life, data=impair)
```

Summary Output

```
> summary(polr(mental ~ ses + life, data = impair))
Coefficients:
          Value Std. Error t value
seshigh -1.1112     0.6109  -1.819
life     0.3189     0.1210   2.635

Intercepts:
                Value   Std. Error t value
none|weak       -0.2819  0.6423    -0.4389
weak|moderate    1.2128  0.6607     1.8357
moderate|strong  2.2094  0.7210     3.0644

Residual Deviance: 99.0979
AIC: 109.0979
```

Applied Statistical Regression

AS 2012 – Week 13

Inference

Again, instead of performing single hypothesis tests, it is better to run deviance tests for nested models.

We first try to exclude predictor `ses`:

```
> fit.life <- polr(mental ~ life, data=impair)
> deviance(fit.life)-deviance(fit)
[1] 3.429180
> pchisq(3.429180, fit$edf-fit.life$edf, lower=FALSE)
[1] 0.0640539
```

→ p-value exceeds 0.05, thus ses is not significant!

Inference

We removed predictor ses from the model, can we also remove the second predictor life? And what kind of model is this?

We now try to exclude predictor `life` from the already reduced model:

```
> fit.empty <- polr(mental ~ 1, data=impair)
> deviance(fit.empty)-deviance(fit.life)
[1] 6.514977
> pchisq(6.514977, fit.life$edf-fit.empty$edf, lower=FALSE)
[1] 0.01069697
```

→ p-value smaller than 0.05, thus `life` is significant!

Applied Statistical Regression

AS 2012 – Week 13

Prediction

As with the multinomial logit model, R allows convenient prediction of either probabilities or class membership. We obtain:

```
> predict(fit.life, type="probs")
        none      weak   moderate    strong
1 0.49337624 0.3037364 0.11173924 0.09114810
2 0.08867378 0.1932184 0.21717188 0.50093592
3 0.29105068 0.3324785 0.18429073 0.19218007
4 0.35380472 0.3345600 0.16025764 0.15137767
5 0.42203463 0.3245379 0.13545441 0.11797305
6 0.56498863 0.2747487 0.09032363 0.06993902
```

→ predicted class is the one with maximal probability

Applied Statistical Regression

AS 2012 – Week 13

Informations on the Exam

- The exam will be on February 11, 2012, 9-11 (provisional)! 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.

Informations on the Exam

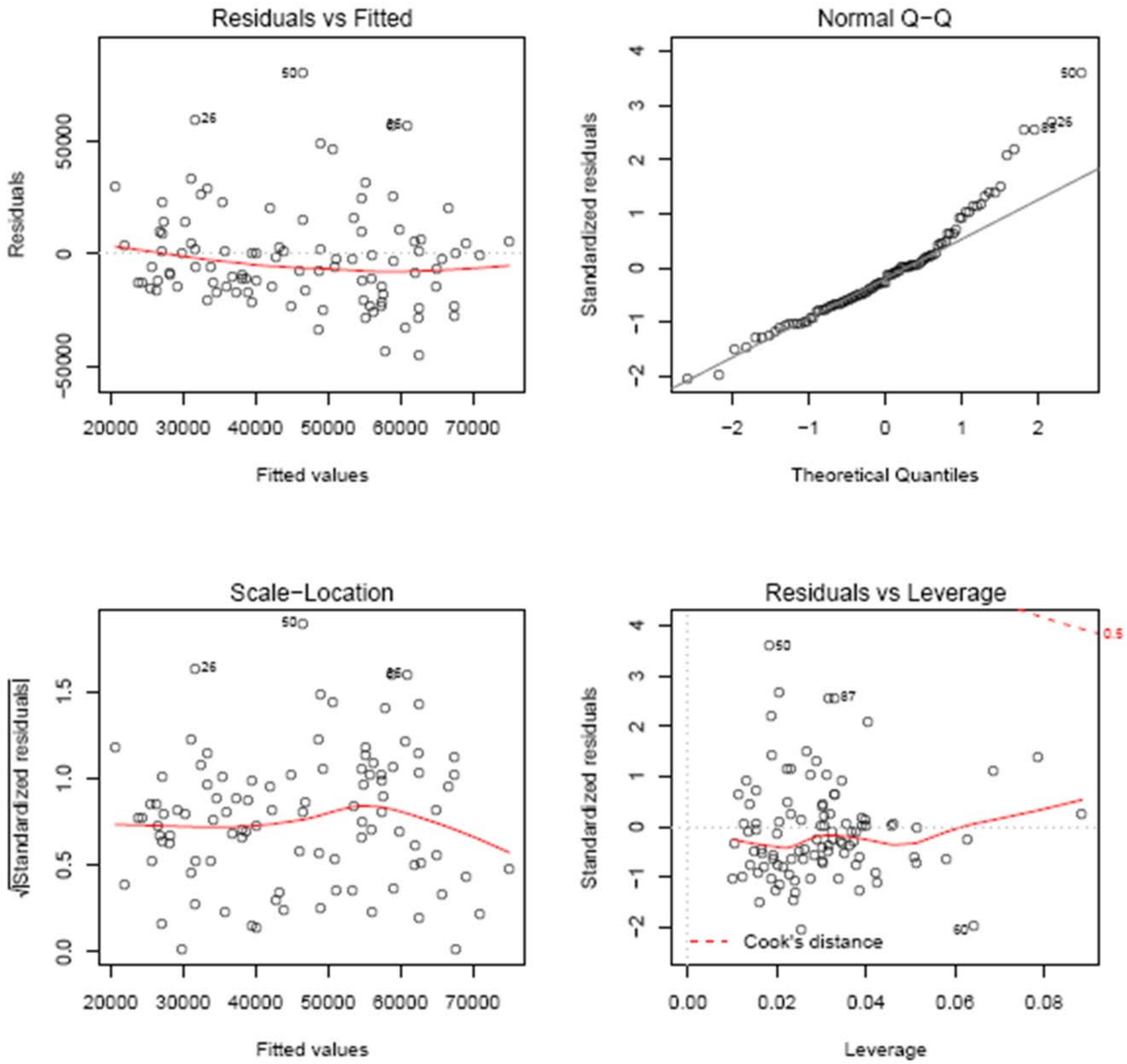
- With the exam, we will try our best to check whether you are proficient in applied regression. This means choosing the right models, interpreting output and suggesting analysis strategies.

- 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.

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

Applied Statistical Regression
AS 2012 – Week 13

Sample Questions from Previous Exams



Applied Statistical Regression

AS 2012 – 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

AS 2012 – 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 fulfilled?

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

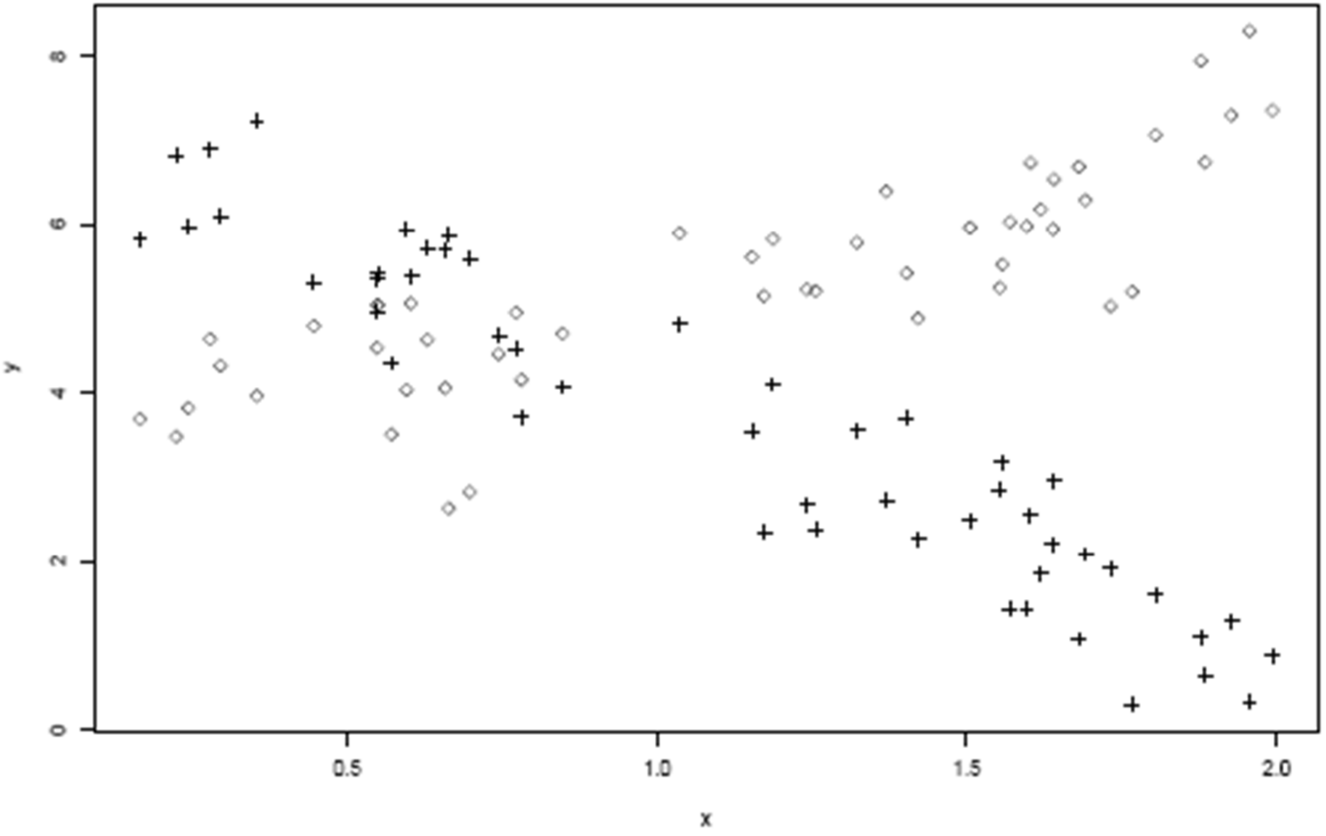
b) If one is only interested in predictions, the model assumptions are not important. These are only important for tests.

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 2012 – Week 13

Sample Questions from Previous Exams



Applied Statistical Regression

AS 2012 – Week 13

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?

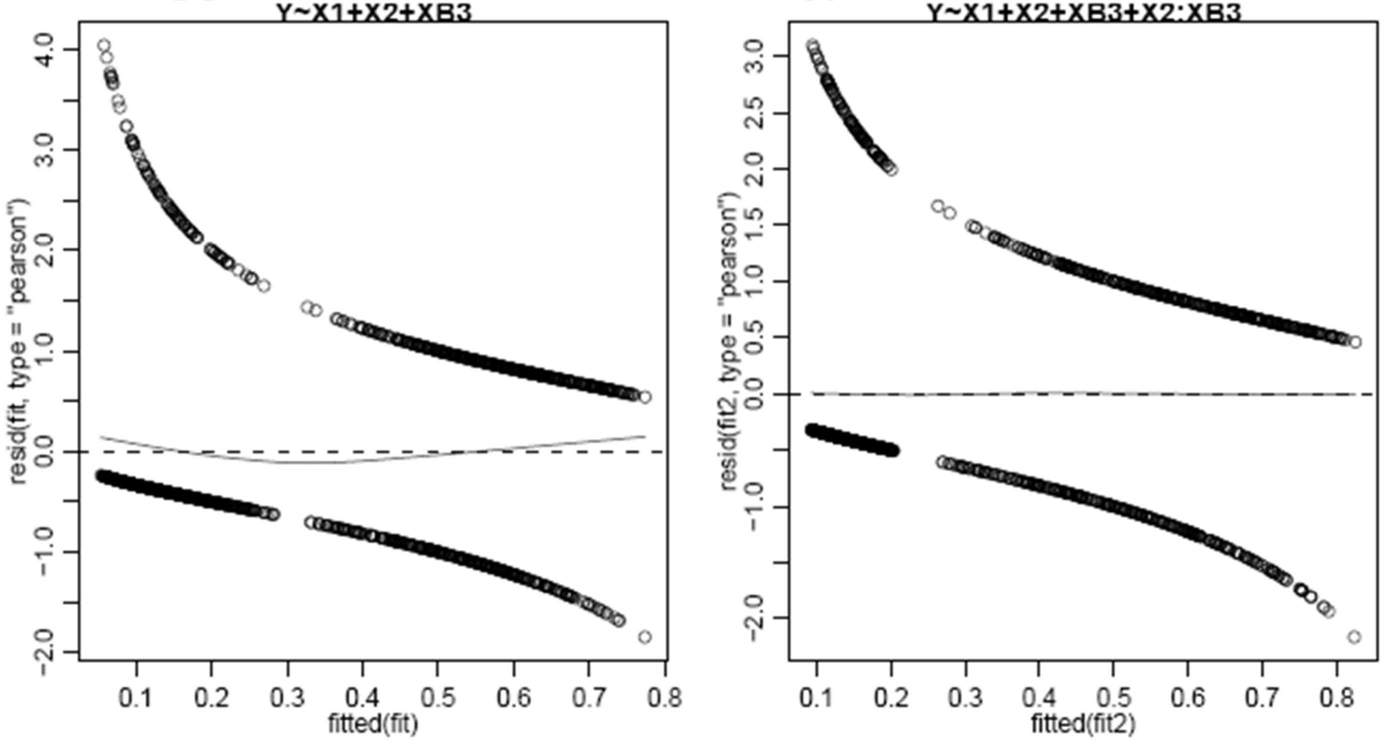
c) Which regression coefficients in your model are clearly positive, cleary negative, approximately 0?

Applied Statistical Regression

AS 2012 – Week 13

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 2012 – Week 13

End of the Course

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