Applied Statistical Regression
HS 2010 – Week 11

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 6, 2010

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

Applied Statistical Regression
HS 2010 – Week 11

Example

American National Election Study 1996: 944 observations

- Response variable: ***party identification***
  - Democrat / Independent / Republican

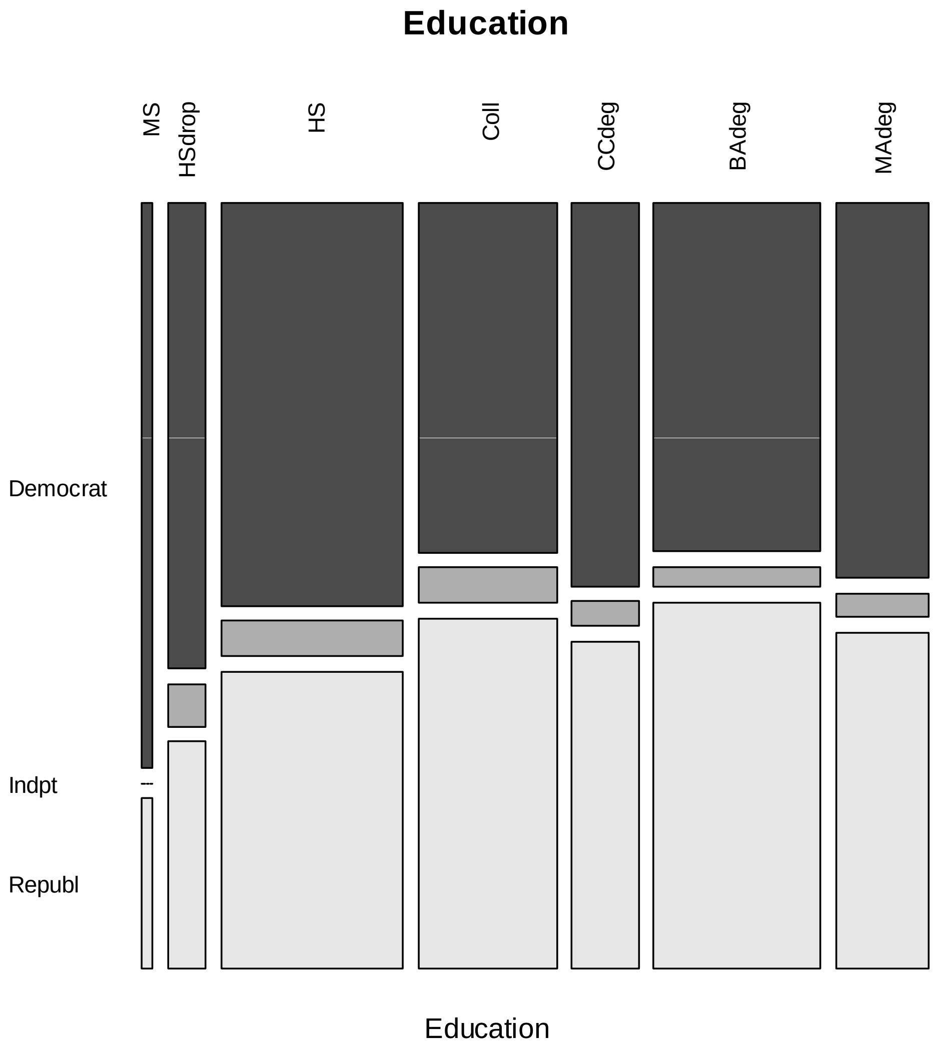
- Predictor 1: ***education***
  - 7 levels: middle school – high school drop - … - MA degree

- Predictor 2: ***income***
  - pseudo-continuous with 24 different values, year income

- Predictor 3: ***age***
  - continuous, age in years

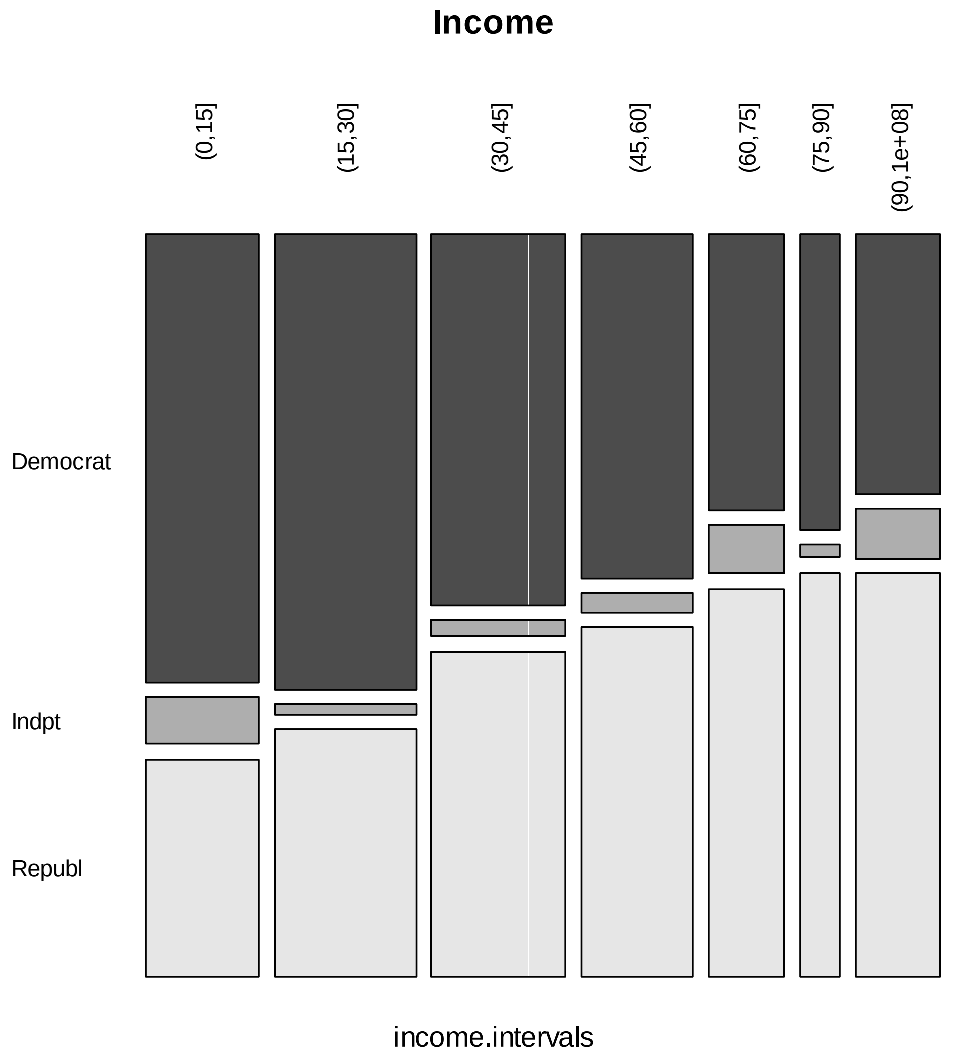
Applied Statistical Regression
HS 2010 – Week 11

Mosaic Plot of Education



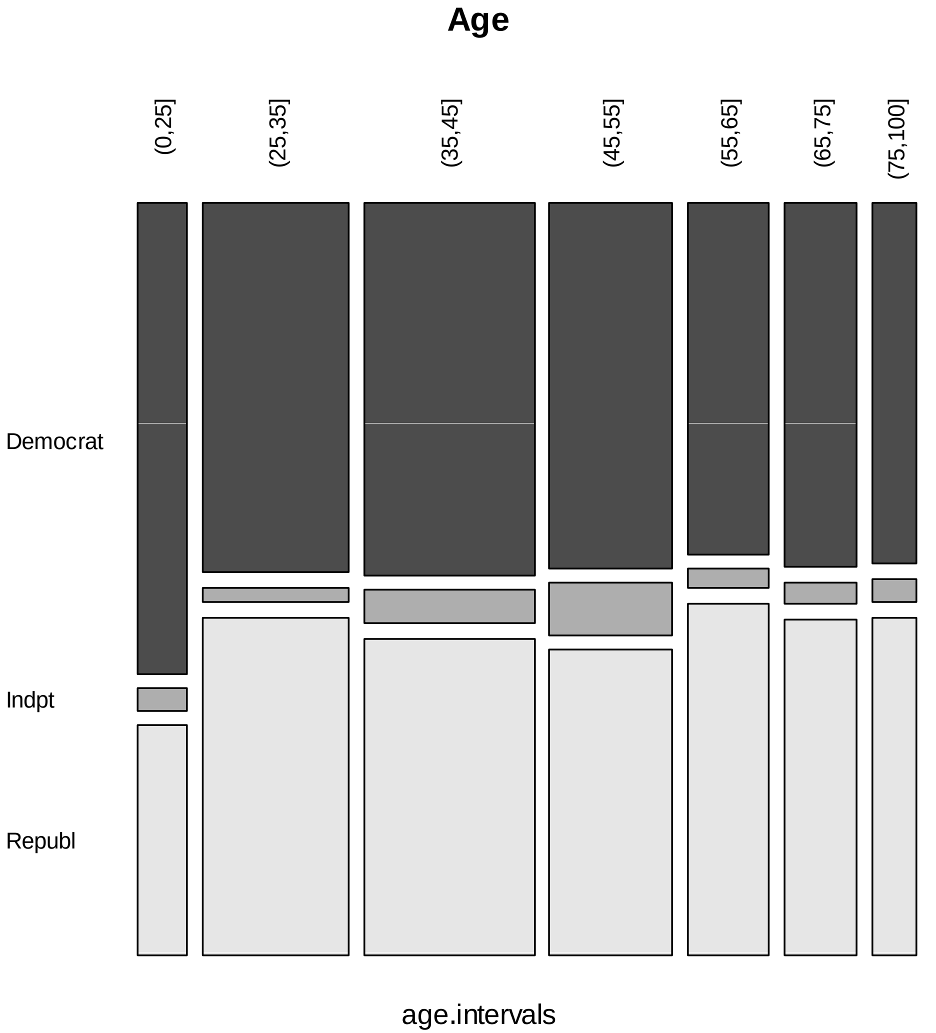
Applied Statistical Regression
HS 2010 – Week 11

Mosaic Plot of Income



Applied Statistical Regression
HS 2010 – Week 11

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
HS 2010 – Week 11

Multinomial Logit Model

- Response $Y_i \in \{1, ..., J\}$
- Ultimate goal: probabilities $p_{ij} = P(Y_i = j)$
- 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
HS 2010 – Week 11

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.

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

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 $\sum_j p_{ij} = 1$.

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!

Applied Statistical Regression
HS 2010 – Week 11

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
```

Zurich University of Applied Sciences

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
HS 2010 – Week 11

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.

Applied Statistical Regression
HS 2010 – Week 11

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
```

- 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
HS 2010 – Week 11

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

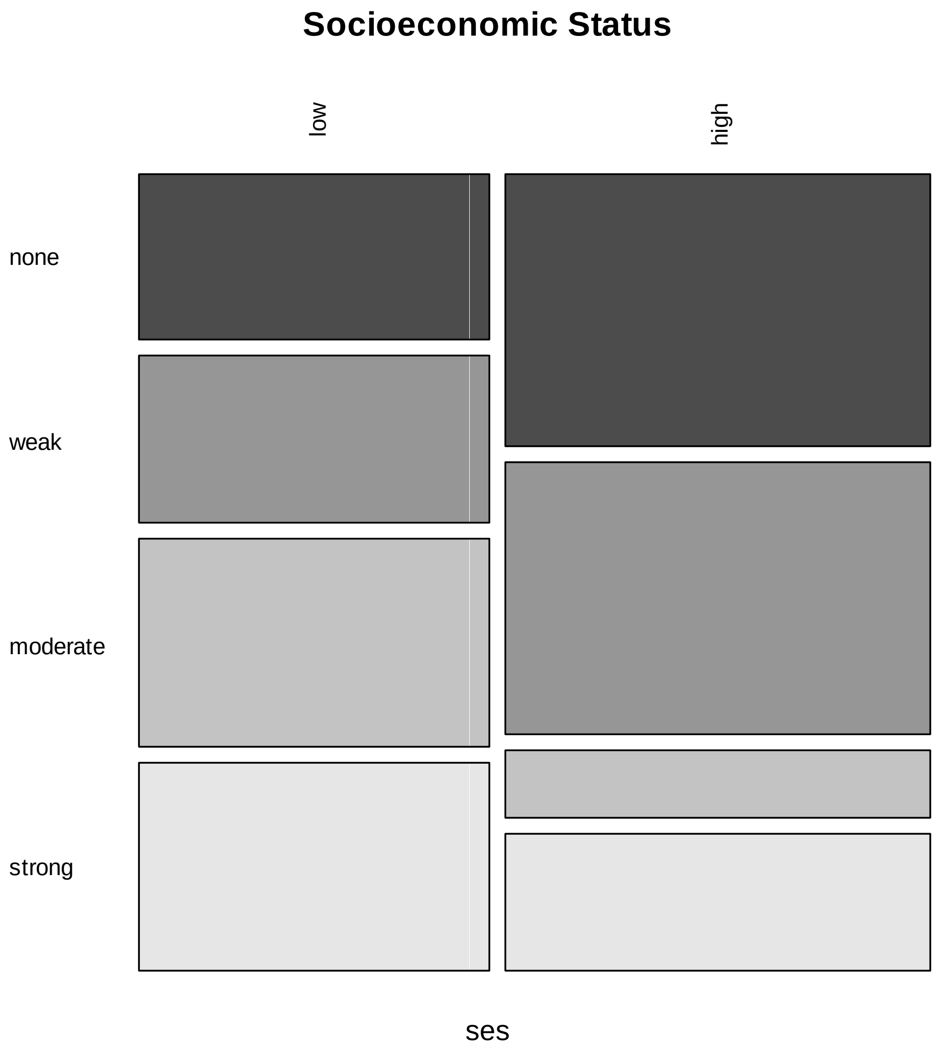
Example

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.

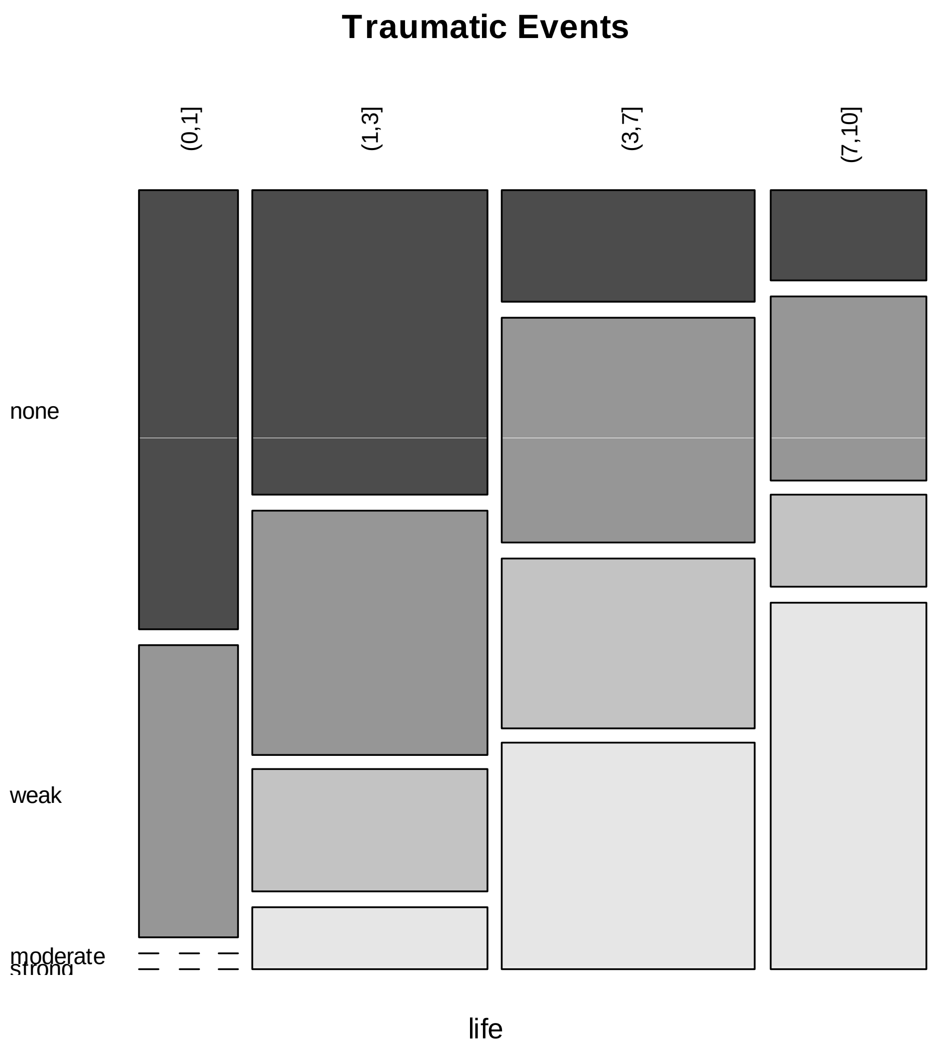
Applied Statistical Regression
HS 2010 – Week 11

Mosaic Plot of SES



Applied Statistical Regression
HS 2010 – Week 11

Mosaic Plot of Life Events



Applied Statistical Regression
HS 2010 – Week 11

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:

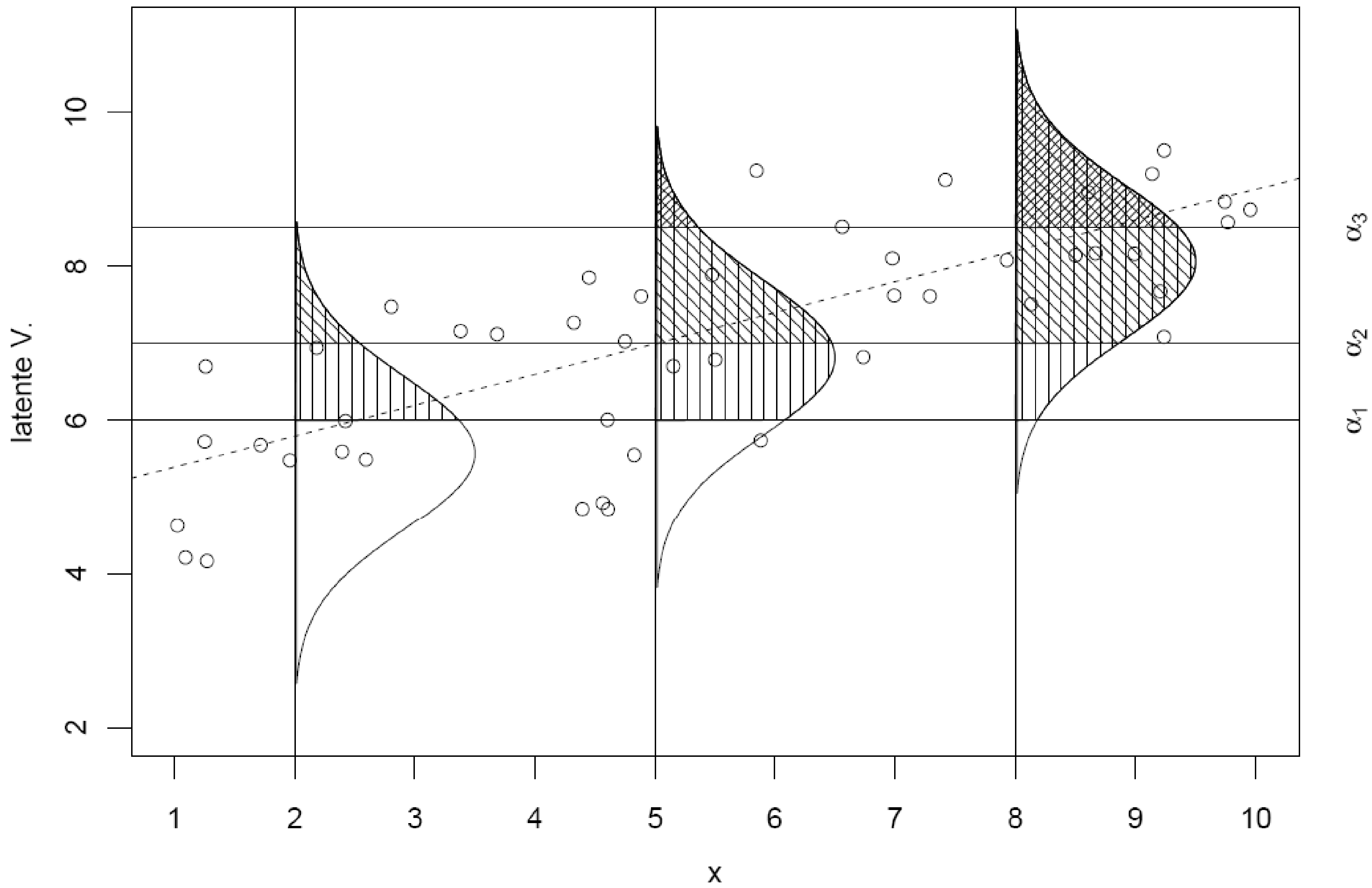
$$Y_i = j \quad \text{if} \quad \alpha_{j-1} < Z_i \leq \alpha_j$$

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
HS 2010 – Week 11

Latent Variable Notion



Proportional Odds Model

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

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

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

Applied Statistical Regression
HS 2010 – Week 11

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)
```

Applied Statistical Regression
HS 2010 – Week 11

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
```

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
HS 2010 – Week 11

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**