

Categorical Regression

Problem 1

a)

Assumptions:

For the multinomial regression we are assuming that the following criterias hold:

- 1) All observations in the data set are independent of each other, that is, the pairs of responses and covariates, $(\mathbf{Y}_i, \mathbf{x}_i)$, are independent.
- 2) Denote the probability of response i being of category r as π_{ir} , that is $P(\mathbf{Y}_i = r) = \pi_{ir}$. Then we assume, when having $c + 1$ categories, that the sum of probabilities equals 1. That is, $\sum_{s=1}^{c+1} \pi_{is} = 1$.
- 3) The covariates are *independent of the response category*, which means that the covariate measurements are the same for all $c + 1$ response categories.

Odds interpretation:

Let c be the number of possible outcomes, p the number of covariates and n the number of observations. This gives the following response matrix \mathbf{y} and design matrix \mathbf{X} .

$$\mathbf{y} = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,c} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,c} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,c} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

In a multinomial GLM, the linear predictor is constructed as

$$\boldsymbol{\eta}_i = \begin{pmatrix} \eta_{i1} \\ \eta_{i2} \\ \vdots \\ \eta_{i,c} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_i^T \boldsymbol{\beta}_1 \\ \mathbf{x}_i^T \boldsymbol{\beta}_2 \\ \vdots \\ \mathbf{x}_i^T \boldsymbol{\beta}_c \end{pmatrix}$$

So we have c different linear predictors for the c response categories, consisting of vectors \mathbf{x}_i^T and $\boldsymbol{\beta}_i$ being the covariate measurements and associated coefficients.

The probability for observations number i to have the outcome r is $\pi_{i,r}$. We let $\pi_{i,r} \propto \exp(\mathbf{x}_i^T \boldsymbol{\beta}_r)$ and normalize, giving

$$\pi_{i,r} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_r)}{\sum_{s=1}^{c+1} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_s)}, \quad \text{for } r = 1, 2, \dots, c+1.$$

But this is unidentifiable, and we therefore introduce a reference category (a dummy variable), which yields

$$\pi_{i,r} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_r)}{1 + \sum_{s=1}^c \exp(\mathbf{x}_i^T \boldsymbol{\beta}_s)}, \quad \text{for } r = 1, 2, \dots, c.$$

If there is a unit change in one of the covariates, for example if the age of person i is changed by one, we denote these two covariate vectors as the original \mathbf{x}_i and \mathbf{x}_i^* .

Let $\pi_{i,r}^*$ denote the probability of a certain outcome when age is increased by 1. Let r_1 and r_2 be two different outcomes, say 'single' and 'widowed'. Now we are interested in how the unit change affected the ratio between

the probabilities for the person being widowed or single. Computing the ratio of the ratios will give us an idea:

$$\frac{\pi_{i,r_1}^* / \pi_{i,r_2}^*}{\pi_{i,r_1} / \pi_{i,r_2}}.$$

Inserting expressions, this yields:

$$\frac{\pi_{i,r_1}^* / \pi_{i,r_2}^*}{\pi_{i,r_1} / \pi_{i,r_2}} = \frac{\exp(x_i^{*T}(\beta_{r_1} - \beta_{r_2}))}{\exp(x_i^T(\beta_{r_1} - \beta_{r_2}))}.$$

Let $\mathbf{x}_i^* = \mathbf{x}_i + [0, 1]^T = [0, x_{i1} + 1]^T$, then the intercept terms cancel, and the unit change in the covariate leads to:

$$\frac{\pi_{i,r_1}^* / \pi_{i,r_2}^*}{\pi_{i,r_1} / \pi_{i,r_2}} = \exp(\beta_{r_1,1} - \beta_{r_2,1})$$

Odds for given category and monotonicity:

For a given category r , the probability of the response variable being of category r , is given as

$$\pi_{ir} = \frac{\exp(\mathbf{x}_i^T \beta_r)}{1 + \sum_{s=1}^c \exp(\mathbf{x}_i^T \beta_s)}$$

the ratio between π_{ir} and its complement, can be expressed as

$$\frac{\pi_{ir}}{1 - \pi_{ir}} = \frac{\frac{\exp(\mathbf{x}_i^T \beta_r)}{1 + \sum_{s=1}^c \exp(\mathbf{x}_i^T \beta_s)}}{1 - \frac{\exp(\mathbf{x}_i^T \beta_r)}{1 + \sum_{s=1}^c \exp(\mathbf{x}_i^T \beta_s)}} = \frac{\exp(\mathbf{x}_i^T \beta_r)}{1 + \sum_{s=1, s \neq r}^c \exp(\mathbf{x}_i^T \beta_s)}$$

Now we can evaluate how the odds changes for a unit increase of age. Again let $\mathbf{x}_i^* = \mathbf{x}_i + [0, 1]^T = [0, x_{i1} + 1]^T$ by looking at the ratio

$$\frac{\pi_{ir}^* / 1 - \pi_{ir}^*}{\pi_{ir} / 1 - \pi_{ir}} = \frac{\frac{\exp(\mathbf{x}_i^{*T} \beta_r)}{1 + \sum_{s=1, s \neq r}^c \exp(\mathbf{x}_i^{*T} \beta_s)}}{\frac{\exp(\mathbf{x}_i^T \beta_r)}{1 + \sum_{s=1, s \neq r}^c \exp(\mathbf{x}_i^T \beta_s)}} = \exp(\beta_{r,1}) \cdot \frac{1 + \sum_{s=1, s \neq r}^c \exp(\mathbf{x}_i^T \beta_s)}{1 + \sum_{s=1, s \neq r}^c \exp(\mathbf{x}_i^{*T} \beta_s)} = \exp(\beta_{r,1}) \cdot \frac{1 - \pi_{ir}}{1 - \pi_{ir}^*}$$

Hence we conclude that the same odds interpretation which applied for $\frac{\pi_{ia}}{\pi_{ia}}$ does *not* hold in the case for $\frac{\pi_{ir}}{1 - \pi_{ir}}$.

If the probabilities of belonging to the different categories would be monotonic functions of age, this would mean that the probabilities either only rise or only decrease as age increases. We would expect this to not hold, as the probability of being married should be increasing for lower values of *age*, and decreasing for higher ages of *age*. People tend to get married in their twenties or thirties, moving from the category ‘single’ to ‘married’. Later on in life, people will lose their significant other and go from ‘married’ to ‘widowed’. This is clearly supported by the plot in **b)**, where the probability of ‘married’ first increases, then decreases.

Test for significance of age

```
mod2 <- vglm(mstatus ~ 1, multinomial)
summary(mod2)
```

```
##
## Call:
## vglm(formula = mstatus ~ 1, family = multinomial)
##
```

```
## Pearson residuals:
##           Min       1Q   Median       3Q      Max
## log(mu[,1]/mu[,4]) -3.435 -0.1199 -0.1199 -0.1199 3.6006
## log(mu[,2]/mu[,4]) -4.957  0.4542  0.4542  0.4542 0.4542
## log(mu[,3]/mu[,4]) -3.906 -0.2149 -0.2149 -0.2149 2.1758
##
## Coefficients:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  1.11014    0.10752   10.32  <2e-16 ***
## (Intercept):2  3.72685    0.09437   39.49  <2e-16 ***
## (Intercept):3  1.95334    0.09964   19.60  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,4]), log(mu[,2]/mu[,4]),
## log(mu[,3]/mu[,4])
##
## Residual deviance: 8423.723 on 18156 degrees of freedom
##
## Log-likelihood: -4211.861 on 18156 degrees of freedom
##
## Number of Fisher scoring iterations: 6
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2'
##
## Reference group is level 4 of the response
dev_diff = deviance(mod2) - deviance(mod1)
df.diff = df.residual(mod2) - df.residual(mod1)
1 - pchisq(dev_diff, df.diff) #significant

## [1] 0
```

We want to test if the linear effect of age is statistically significant by examining the deviance residuals. We compute two models, one with age as a covariate and one without, respectively m_{age} and m_1 .

We compute the residual deviance (found in the summary) and the difference in degrees of freedom. Then we get the value 1600.933 with 3 degrees of freedom, and the probability for observing a value such as this or more extreme is a hard 0. Thus, we can safely conclude that the linear effect of age is statistically significant.

b

```
test = round(predict(mod1, newdata = data.frame("age" = seq(16,88)), type = c("response")),3)
test = data.frame(test,row.names=seq(16,88))

Aic <- round(extractAIC(mod1)[2])

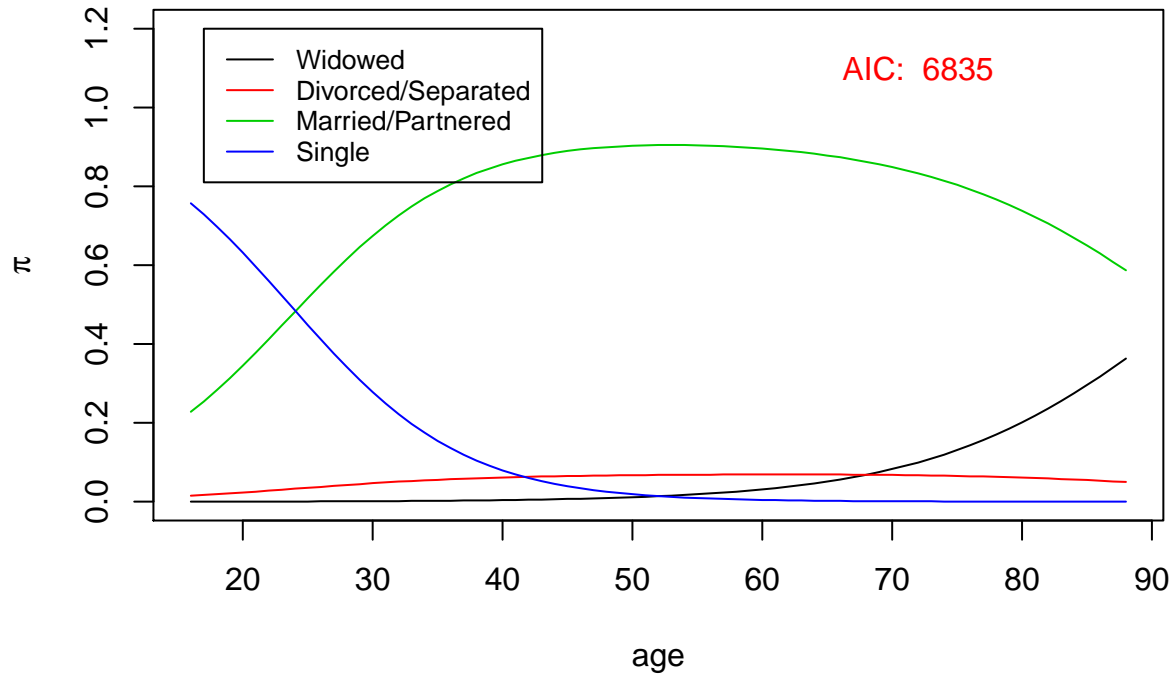
plot(rownames(test),test$Widowed, ylim = c(0,1.2), type = 'line', xlab='age', ylab = expression(pi))

## Warning in plot.xy(xy, type, ...): plot type 'line' will be truncated to
## first character
```

```

text(x=72,y=1.1,paste('AIC: ',toString(Aic)),col = 2)
lines(rownames(test),test$Divorced.Separated, col=2)
lines(rownames(test),test$Married.Partnered, col=3)
lines(rownames(test),test$Single, col=4)
legend(17, 1.2, legend=c("Widowed", "Divorced/Separated", "Married/Partnered", "Single"), col=c(1, 2, 3, 4))

```



In the figure above the probabilities for different relationship statuses are plotted as a function of age. The probability of being single is very high when age is around 25 or less, and decreases rapidly until the age nears 50. Until the age of 50 the sum of the probabilities for being married and single seem to sum to 1, with the probability of being divorced/separated quite low for all ages. When age grows beyond 50, the probability of being widowed increases while the probability of being Married decreases. In conclusion, the graphs look pretty much as expected.

c

You can also embed plots, for example:

```

mod2pol <- vglm(mstatus ~ poly(age,2), multinomial)
mod3pol <- vglm(mstatus ~ poly(age,3), multinomial)
mod4pol <- vglm(mstatus ~ poly(age,4), multinomial)
mod5pol <- vglm(mstatus ~ poly(age,5), multinomial)
prediction2pol <- data.frame(predict(mod2pol,type = "response", newdata = data.frame("age" = seq(16,88)))
prediction3pol <-data.frame(predict(mod3pol,type = "response", newdata = data.frame("age" = seq(16,88)))
prediction4pol <- data.frame(predict(mod4pol,type = "response", newdata = data.frame("age" = seq(16,88)))
prediction5pol <-data.frame(predict(mod5pol,type = "response", newdata = data.frame("age" = seq(16,88)))

Aic2 <- round(extractAIC(mod2pol)[2])
Aic3 = round(extractAIC(mod3pol)[2])
Aic4 =round(extractAIC(mod4pol)[2])
Aic5 =round(extractAIC(mod5pol)[2])

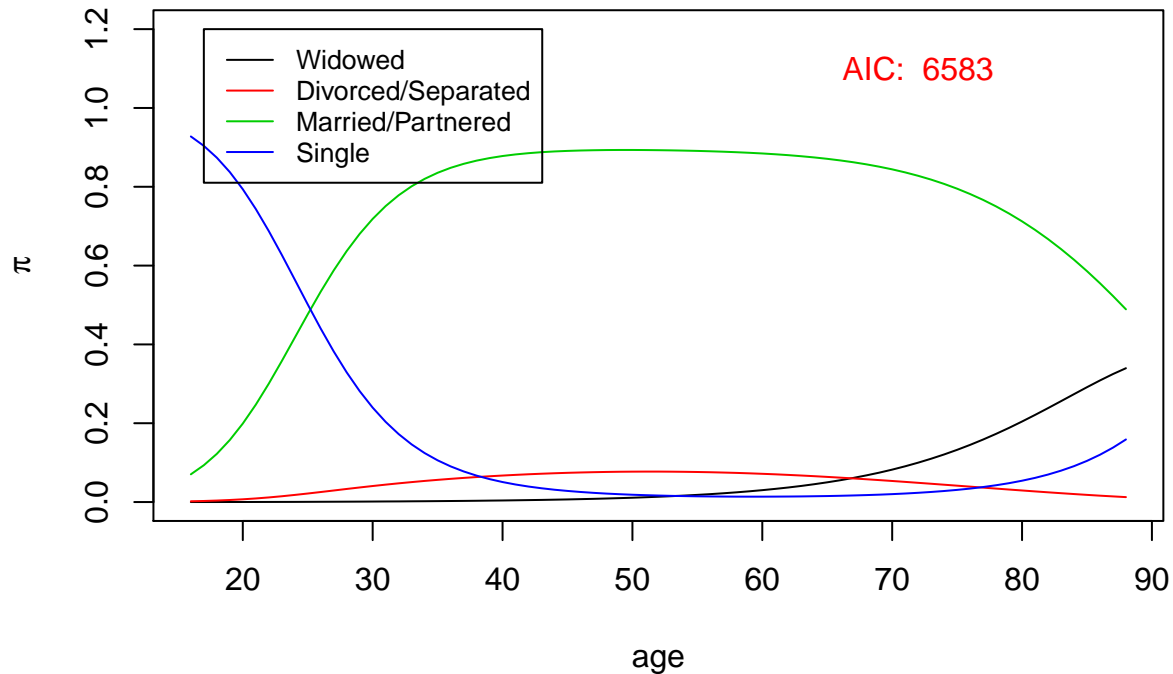
plot(rownames(prediction2pol),prediction2pol$Widowed, ylim = c(0,1.2), main='2th power regression', type='l')

```

```
## Warning in plot.xy(xy, type, ...): plot type 'line' will be truncated to
## first character
```

```
text(x=72,y=1.1,paste('AIC: ',toString(Aic2)),col = 2)
lines(rownames(prediction2pol),prediction2pol$Divorced.Separated, col=2)
lines(rownames(prediction2pol),prediction2pol$Married.Partnered, col=3)
lines(rownames(prediction2pol),prediction2pol$Single, col=4)
legend(17, 1.2, legend=c("Widowed", "Divorced/Separated", "Married/Partnered", "Single"), col=c(1, 2, 3, 4))
```

2th power regression

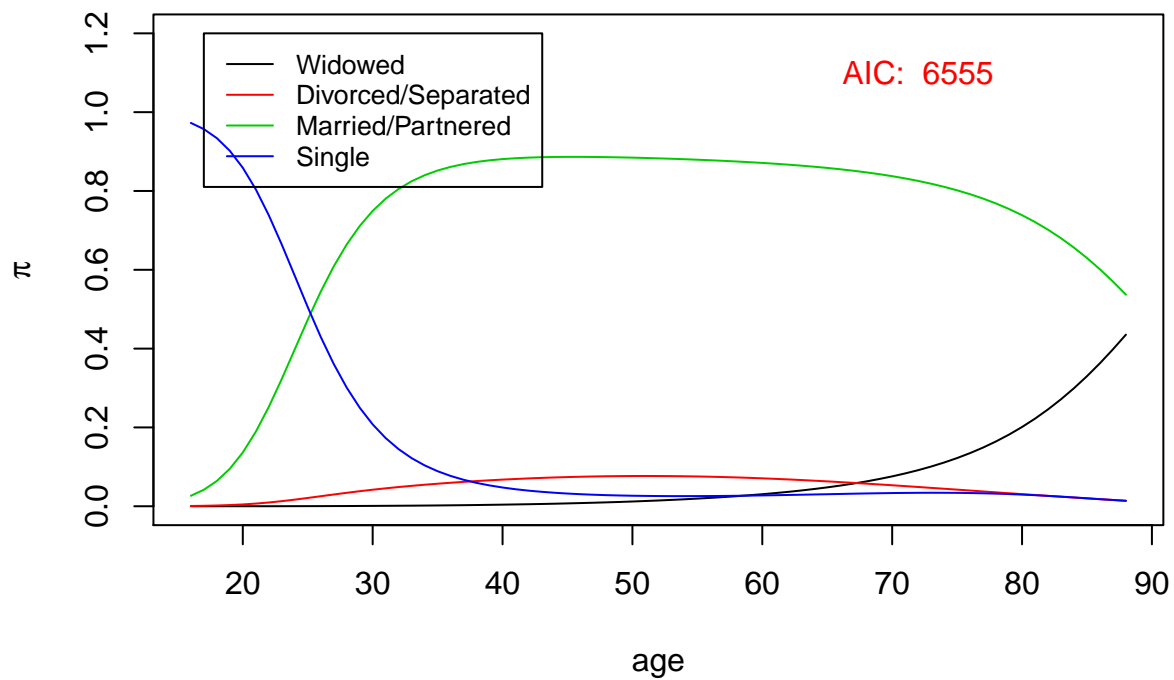


```
plot(rownames(prediction3pol),prediction3pol$Widowed, ylim = c(0,1.2), main='3th power regression', type='l')
```

```
## Warning in plot.xy(xy, type, ...): plot type 'line' will be truncated to
## first character
```

```
lines(rownames(prediction3pol),prediction3pol$Divorced.Separated, col=2)
text(x=72,y=1.1,paste('AIC: ',toString(Aic3)),col = 2)
lines(rownames(prediction3pol),prediction3pol$Married.Partnered, col=3)
lines(rownames(prediction3pol),prediction3pol$Single, col=4)
legend(17, 1.2, legend=c("Widowed", "Divorced/Separated", "Married/Partnered", "Single"), col=c(1, 2, 3, 4))
```

3th power regression

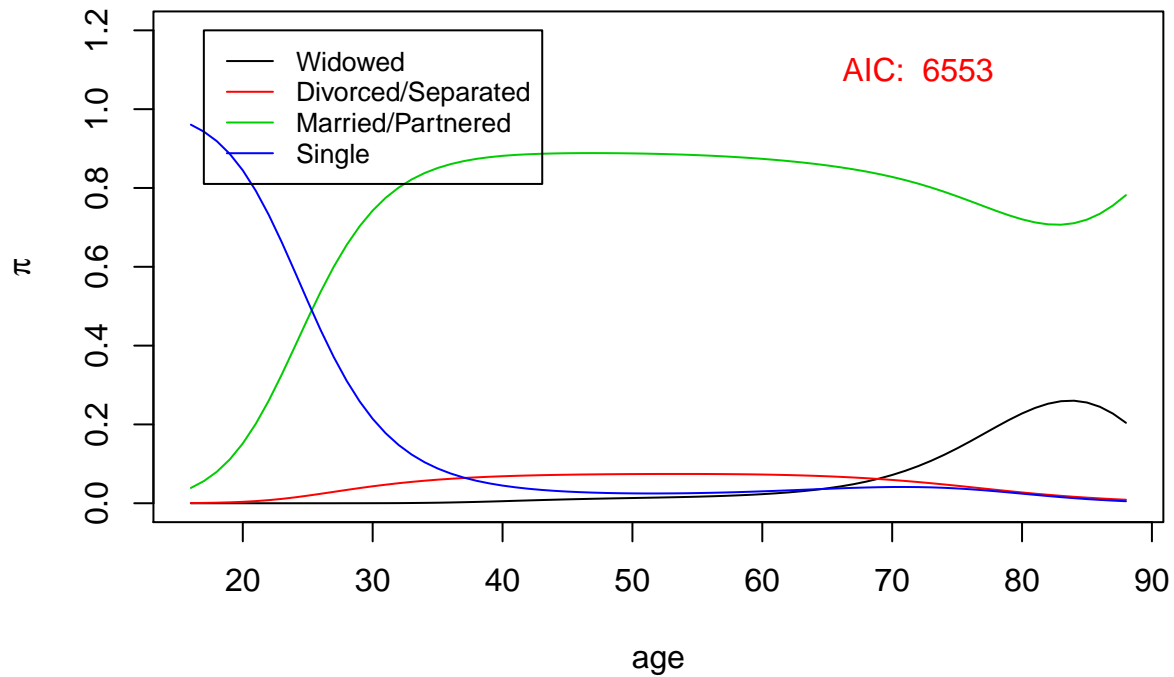


```
plot(rownames(prediction4pol),prediction4pol$Widowed, ylim = c(0,1.2), main='4th power regression', type='l')
```

```
## Warning in plot.xy(xy, type, ...): plot type 'line' will be truncated to
## first character
```

```
lines(rownames(prediction4pol),prediction4pol$Divorced.Separated, col=2)
text(x=72,y=1.1,paste('AIC: ',toString(Aic4)),col = 2)
lines(rownames(prediction4pol),prediction4pol$Married.Partnered, col=3)
lines(rownames(prediction4pol),prediction4pol$Single, col=4)
legend(17, 1.2, legend=c("Widowed", "Divorced/Separated", "Married/Partnered", "Single"), col=c(1, 2, 3, 4))
```

4th power regression

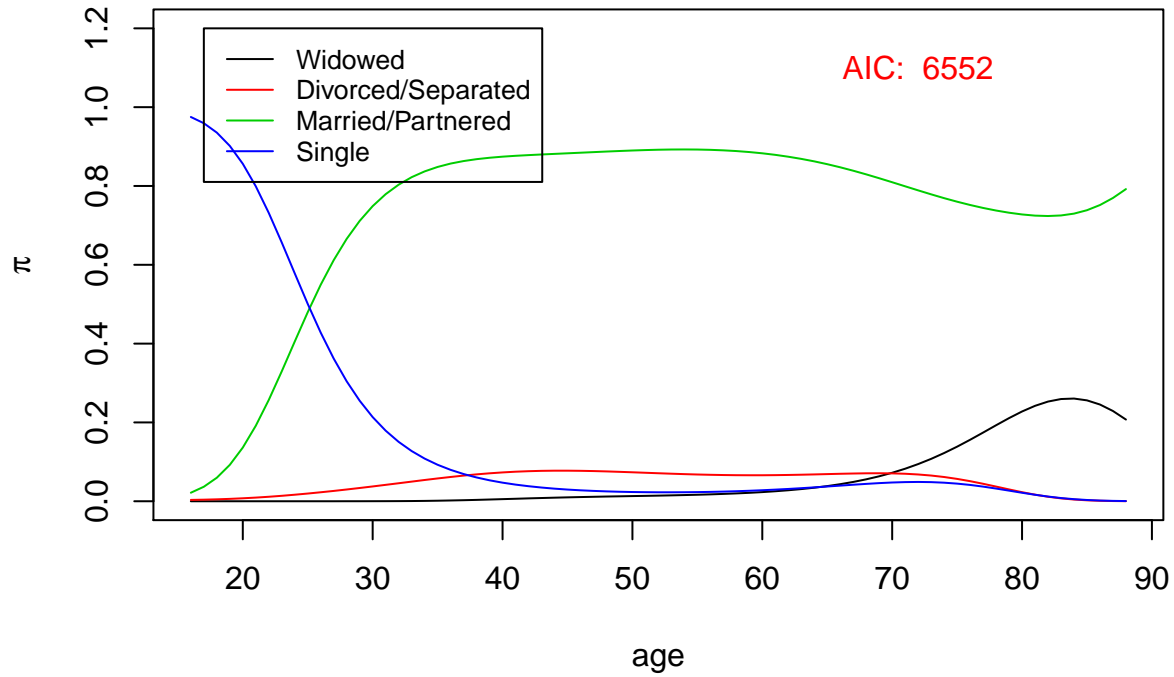


```
plot(rownames(prediction5pol),prediction5pol$Widowed, ylim = c(0,1.2), main='5th power regression', type='l')
```

```
## Warning in plot.xy(xy, type, ...): plot type 'line' will be truncated to
## first character
```

```
lines(rownames(prediction5pol),prediction5pol$Divorced.Separated, col=2)
text(x=72,y=1.1,paste('AIC: ',toString(Aic5)),col = 2)
lines(rownames(prediction5pol),prediction5pol$Married.Partnered, col=3)
lines(rownames(prediction5pol),prediction5pol$Single, col=4)
legend(17, 1.2, legend=c("Widowed", "Divorced/Separated", "Married/Partnered", "Single"), col=c(1, 2, 3, 4))
```

5th power regression



By comparing AIC values we see that the AIC-value decreases as the order increases. The fifth order model, denoted by $m_{o=5}$ has the lowest AIC-value, while $m_{o=4}$ and $m_{o=3}$ are also decent candidates. The rest have a significantly lower AIC.

At first glance, it seems contra-intuitive that $m_{o=5}$ and $m_{o=4}$ are deemed good, because they seem pretty funky for $age > 85$. It says that for persons older than 85 years, the rate of widowed people suddenly drops, while the rate of married people increases. This probably is not the case, and thus one could argue that $m_{o=3}$ may be a better choice.

The reason that this funky behaviour does not affect the AIC significantly, is the very small number of people that actually live to be 85 years or more. Thus categorizing them wrongly does not reduce the log likelihood by much and in turn the AIC does not suffer.