GLM Project 2

TMA4315 Generalized linear models H2020

magnwoln@stud.ntnu.no, siljeanf@stud.ntnu.no, 10020, 10013

06 November, 2020

Problem a

In this model the response variable mstatus is given as one out of four categories: Divorced/Separated, Married/Partnered, Single, Widowed. The last category Widowed is the reference category. We fit a multinomial regression model with one covariate age.

```
attach(marital.nz)
#Fitting a multinomial regression with linear effect of age
mod1 <- vglm(mstatus ~ age, family = 'multinomial')
summary(mod1)
```

```
##
## Call:
## vglm(formula = mstatus ~ age, family = "multinomial")
##
## Pearson residuals:
##                         Min      1Q   Median       3Q    Max
## log(mu[,1]/mu[,4]) -11.75 -0.1441 -0.13965 -0.13372  5.706
## log(mu[,2]/mu[,4]) -13.53  0.2871  0.31147  0.40939  1.212
## log(mu[,3]/mu[,4]) -12.47 -0.2364 -0.09098 -0.02037 82.311
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  6.753157   0.515150   13.11   <2e-16 ***
## (Intercept):2  9.531824   0.482073   19.77   <2e-16 ***
## (Intercept):3 13.121214   0.513771   25.54   <2e-16 ***
## age:1         -0.099335   0.008043  -12.35   <2e-16 ***
## age:2         -0.102873   0.007100  -14.49   <2e-16 ***
## age:3         -0.252080   0.008955  -28.15   <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: 6822.79 on 18153 degrees of freedom
##
## Log-likelihood: -3411.395 on 18153 degrees of freedom
##
## Number of Fisher scoring iterations: 7
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
```

```
## '(Intercept):2', 'age:3'
##
##
## Reference group is level  4  of the response
```

Part1: Asumptions The categorical response variable $Y_i \in \{1, 2, 3, 4\}$ is measured on a nominal scale. The covariates $\mathbf{x}_i$, which consists of only the intercept and age are assumed to be independent of the response category, meaning the martial status and the age of each observation/person used in this experiment is independent of the rest. We also assume that each observation fits into exactly one of the four categories of the response, martial status. The probability of occurence for category r, where $r = 1, 2, 3$ and i denotes observation number, is given as

$$P(Y_i = r) = \pi_{ir} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_r)}{1 + \sum_{s=1}^{3} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_s)}$$

The reference category is given as

$$P(Y_i = 4) = P(\text{Widowed}) = \pi_{i,4} = 1 - \pi_{i1} - \pi_{i2} - \pi_{i3} = \frac{1}{1 + \sum_{s=1}^{3} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_s)}$$

We assume that the probability of occurence of a given category is the same for each of the observations i.

Part 2: Odds ratios The parameter vectors for the model with a linear effect of age is $\boldsymbol{\beta}_r = (\beta_{r0}, \beta_{r1})$ where $r = 1, 2, 3$. The probabilities above can also be given as log-odds ratios for $r = 1, 2, 3$,

$$\ln\left(\frac{P(Y_i = r)}{P(Y_i = 4)}\right) = \ln\left(\frac{\pi_{ir}}{\pi_{i,4}}\right) = \eta_{ir} = \mathbf{x}_i^T \boldsymbol{\beta}_r = \beta_{r0} + \beta_{r1}\text{age}$$

where η_{ir} is the linear predictor and the log-odds is the link function, or equivalently

$$\frac{P(Y_i = r)}{P(Y_i = 4)} = \frac{\pi_{ir}}{\pi_{i,4}} = \exp(\eta_{ir}) = \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}_r\right) = \exp(\beta_{r0} + \beta_{r1}\text{age}) = \exp(\beta_{r0}) \exp(\beta_{r1}\text{age}).$$

We can interpret $\exp(\beta_{r1})$ as the increase in relative risk (odds ratio) if the covariate age is increased with one unit, meaning one year increase. Let π_{ir}^* denote the probability for occurence of category r for observation i with one unit change of the covariate age, and $x_i^* = x_i + 1$, then

$$\frac{P(Y_i = r|\text{age} = x_i + 1)/P(Y_i = 4|\text{age} = x_i + 1)}{P(Y_i = r|\text{age} = x_i)/P(Y_i = 4|\text{age} = x_i)} = \frac{\pi_{ir}^*/\pi_{i4}^*}{\pi_{ir}/\pi_{i4}} = \frac{e^{\mathbf{x}_i^{*T}\boldsymbol{\beta}_{r1}}}{e^{\mathbf{x}_i^T\boldsymbol{\beta}_{r1}}} = e^{(\mathbf{x}_i^* - \mathbf{x}_i)^T\boldsymbol{\beta}_{r1}} = e^{(\mathbf{x}_i + 1 - \mathbf{x}_i)^T\boldsymbol{\beta}_{r1}} = e^{\boldsymbol{\beta}_{r1}}$$

```
coef(mod1)
```

```
## (Intercept):1 (Intercept):2 (Intercept):3         age:1         age:2
##    6.75315744    9.53182379   13.12121378   -0.09933456   -0.10287286
##         age:3
##   -0.25207998
```

So if the parameter β_{r1} for example is positive it means that the odds for category r increase only relative to the reference category, which in this project is Widowed. From the summary of the model fitted in 1a part1 we see that the coefficients for age with each of the three categories, $\beta_{11}, \beta_{21}, \beta_{31}$ are negative. This means that the odds for being either Divorced/Separated, Married/Partnered and Single is decreasing relative to being Widowed for an increasing age.

Part 3:

Using the same procedure as in part 2 we calculate the ratio of being and not being in a certain martial status for one unit increase of x_i. We can think of it as looking at the problem as a binary regression problem with only two categories.

$$\frac{\pi^*_{ir}/(1-\pi^*_{ir})}{\pi_{ir}/(1-\pi_{ir})} = \frac{e^{\mathbf{x}_i^{*T}\boldsymbol{\beta}_r}/1+\sum_{s=1,s\neq r}^{3} e^{\mathbf{x}_i^{*T}\boldsymbol{\beta}_s}}{e^{\mathbf{x}_i^{T}\boldsymbol{\beta}_r}/1+\sum_{s=1,s\neq r}^{3} e^{\mathbf{x}_i^{T}\boldsymbol{\beta}_s}} = e^{(\mathbf{x}_i^{*}-\mathbf{x}_i)^{T}\boldsymbol{\beta}_r}\frac{1+\sum_{s=1,s\neq r}^{3} e^{\mathbf{x}_i^{T}\boldsymbol{\beta}_s}}{1+\sum_{s=1,s\neq r}^{3} e^{\mathbf{x}_i^{*T}\boldsymbol{\beta}_s}} = e^{\boldsymbol{\beta}_r}\frac{1+\sum_{s=1,s\neq r}^{3} e^{\mathbf{x}_i^{T}\boldsymbol{\beta}_s}}{1+\sum_{s=1,s\neq r}^{3} e^{\boldsymbol{\beta}_s+\mathbf{x}_i^{T}\boldsymbol{\beta}_s}}$$

Notice that the above expression is dependent on each of the categorical coefficients in the model. In part 2 we found that the odds ratio for one unit increase of age is dependent only on one of these coefficients, β_r. Therefore the same interpretation as used in part 2 does not apply for the above expression.

In order to check if the probabilities of belonging to the different categories π_r are monotonic functions of age we take a look at the binomial interpretation of the odds ratio $\pi_{ir}/(1-\pi_{ir})$. We notice that when changing the reference category in the model such that the coefficients for each category also changes the odds ratio interpretation will stay the same. Meaning the probabilities π_r are independent of choice of reference category. Hence the probabilities of belonging to different categories are not monotonic functions of age.

Part 4: Test linear effect of age In order to test significance of linear effect of age we will perform the following hypothesis test, $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$. We calculate the two sided p-value under H_0 by anova.

```
mod0 <- vglm(mstatus ~ 1, multinomial) #Model only based on intercept
anova(mod0, mod1, test="LRT", type="1") #Comparing the two different models
```

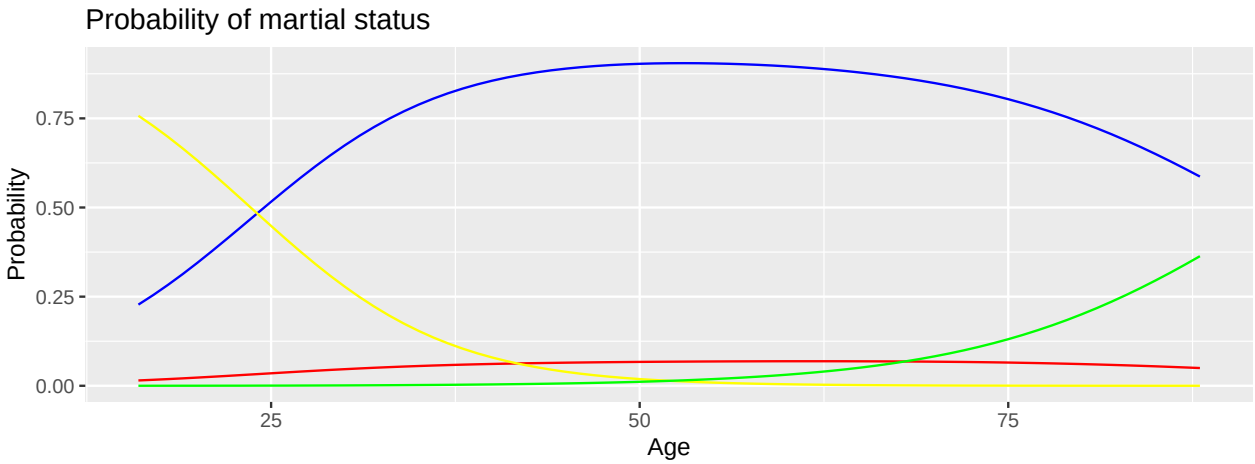
```
## Analysis of Deviance Table
##
## Model 1: mstatus ~ 1
## Model 2: mstatus ~ age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     18156     8423.7
## 2     18153     6822.8  3   1600.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the print-out above we notice that the p-value is smaller than $\alpha = 0.05$ meaning we can reject the null hypothesis saying that the coefficient for age is insignificance. In other words, the we will include linear effect of age.

Problem b

Part 1:

```
M <- data.frame(age = c(16:88)) #Create the list of ages from 16 to 88
prob <- predict(mod1, newdata = M, type = 'response') #Predict the martial status
datapred <- cbind(M, prob) #Create data matrix of age and probabilities
```



In this plot we have plotted the probability for being in the different categories as a function age. The different categories are presented with different colours. The red line represent divorced/seperated, the blue line represent married/partnered, the yellow line represent single and the green line represent widowed.

Part 2:

From the plot we observe that the probability of being married (blue line) is non-monotonus as it is increasing for increasing age until it start to decrease for an age of about 55 years. This seems reasonable as the probability of being divorced (red line) together with the probability of being Widowed (green line) increases for an increasing age. The probability of being single (yellow line) decreases for an increasing age which is reasonable as people get married when getting older.

This plot confirms the conclusion in Problem a, part 3, the probabilities of belonging to the different categories π_r are not monotonic functions of age.

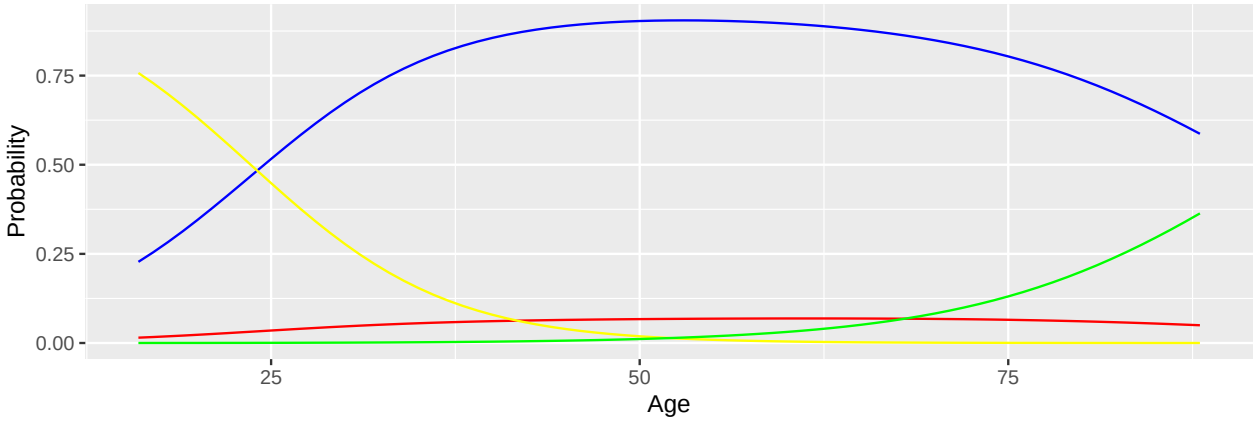
Problem c

Part 1)

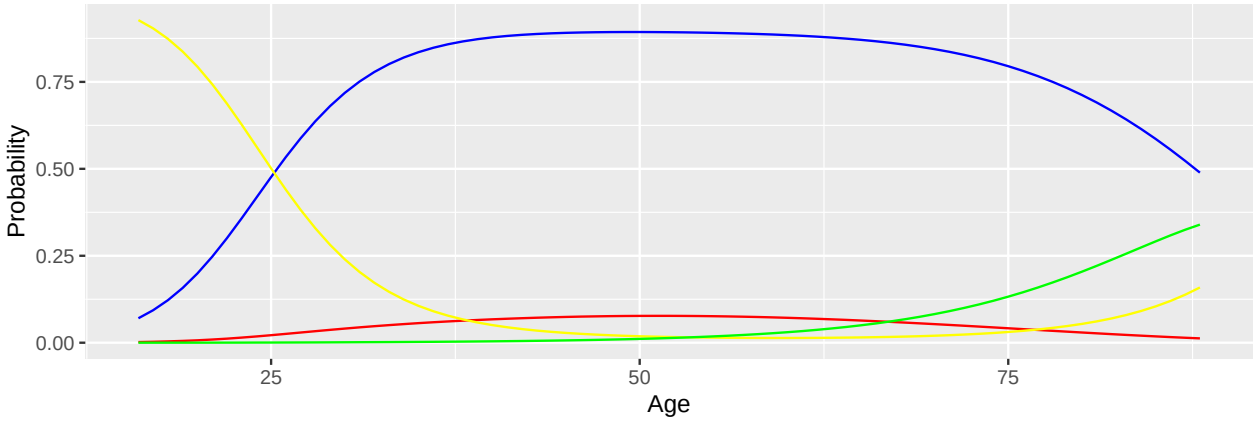
In this task we have tried to improve the model with polyomial regression by includig different powers of age. We use the same plot as in b) with blue representing married/partnered, red representing divorced/seperated, yellow representing single and green representing widowed.

```
deg = c(1,2,3,4,5) #Creatig a list for different degrees
AIClist = c() #Empty list for AIC values

for (d in deg) {
  attach(marital.nz)

  mod <- vglm(mstatus ~ poly(age,d), family = 'multinomial') #Model with polyomial regression with degr

  M <- data.frame(age = c(16:88))

  prob <- predict(mod, newdata = M, type = 'response')

  datapred <- cbind(M, prob)

  print(ggplot(data = datapred, aes(x = datapred$age, y = datapred$`Divorced/Separated`)) + geom_line(da

  AIClist[d] = AIC(mod) #Calculate the AIC for the model and adding it to the list
}
```

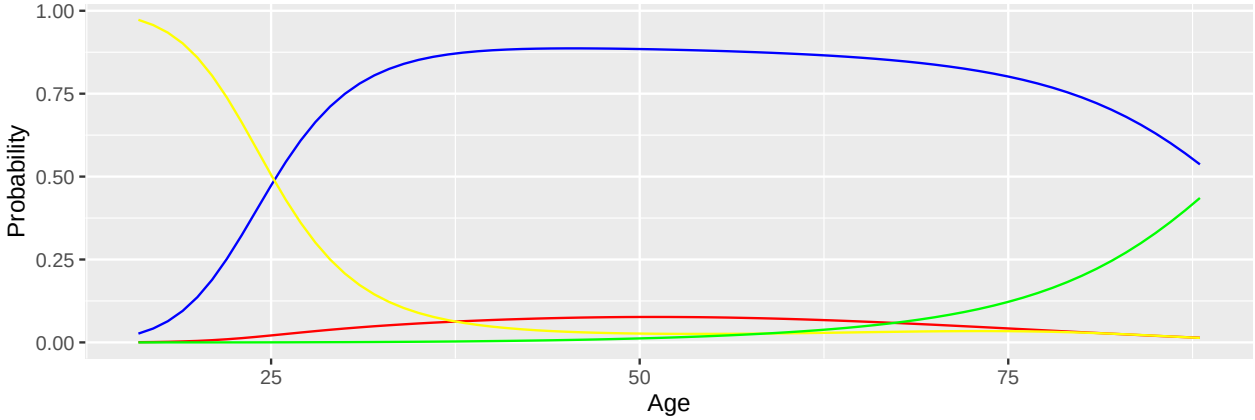
P(martial status) with age to the power of 1

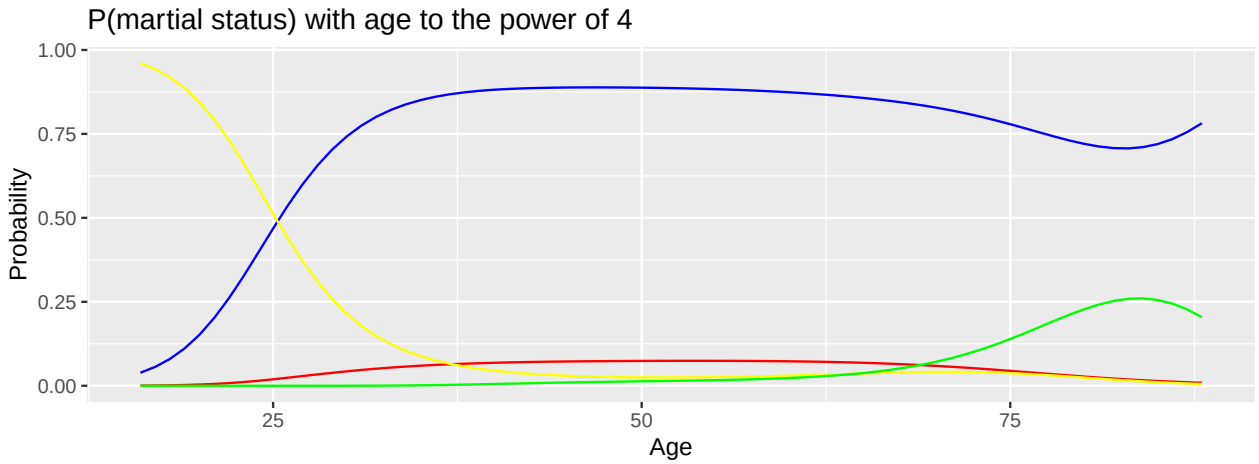


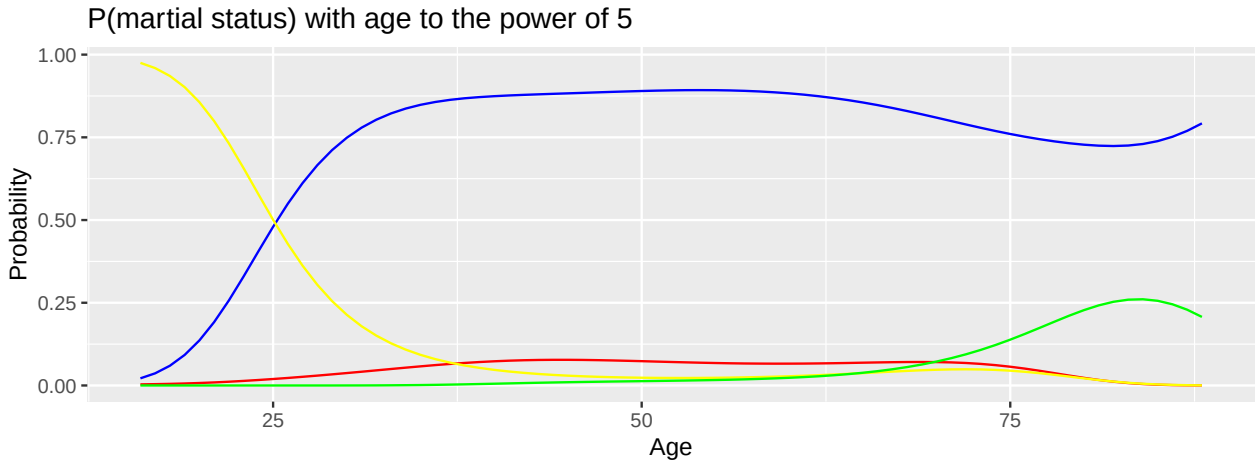
P(martial status) with age to the power of 2



P(martial status) with age to the power of 3



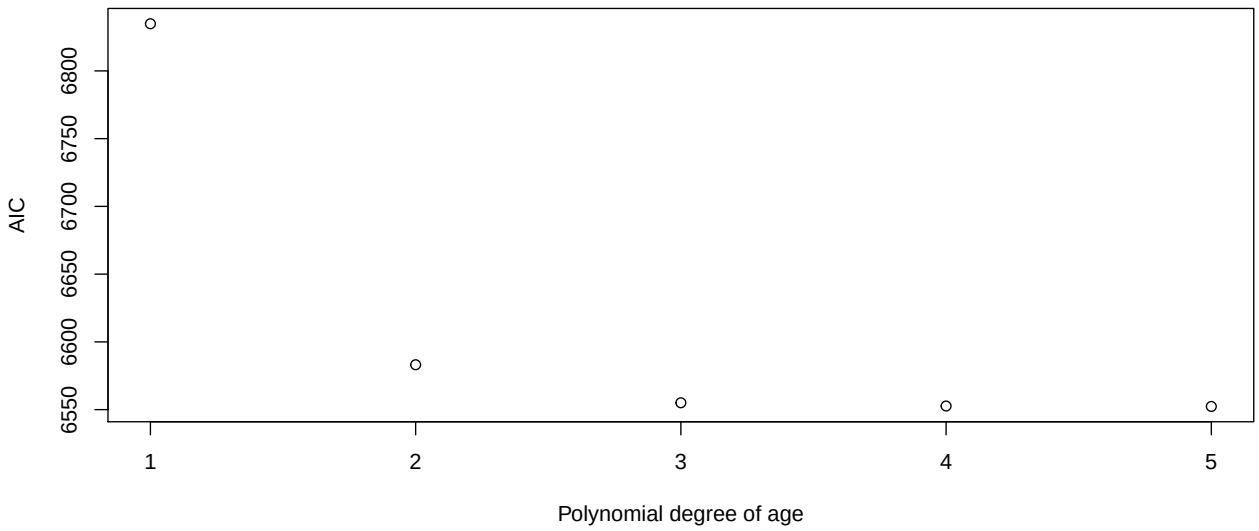





```
print(AIClist)
```

```
## [1] 6834.790 6583.123 6555.048 6552.665 6552.341
```

```
plot(AIClist, xlab = "Polynomial degree of age", ylab = "AIC")
```



Part 2)

AIC is helpfull critera in model selection as it quantifies the loss of information from the following model and as well as penalizing many parameters.

The formula for AIC is

$$AIC = -2l(\hat{\beta}) + 2p$$

where $l(\hat{\beta})$ is the log-likelihood and p is the number of estimated parameters in the model. Here we can see that the last term punishes large models (models with many covariates). The AIC will decrease for models that provides as much information about the data as possible and at the same time includes few independent variables. This is also illustrated from the plot of AIC as a function of polynomial degrees.

Part 3)

Now, do model selection based on the AIC criteria for each of the polynomial regression models in part 1. Based in the AIC criteria showed in the plot in part 1 the "best model" is the polynomial regression including powers of age up to 5. Looking at the plot of probabilities of this model we clearly see that it is not reasonable. For example the probability of being married (blue line) is increasing from age 80 and up, which makes no sense in the real world. Also the probability of being widowed (green line) is not likely to decrease after age 80. Although it is possible to remarry at age 80, it is not common and therefore the model is not reasonable.